Report as of FY2010 for 2008OH64B: "Competitive Learning to Develop a Biomarker Forecasting Tool for Classifying Recreational Water Quality"

Publications

- Dissertations:
    - Motamarri, S. (2010) “Development of a Neural Based Biomarker Forecasting Tool to Classify Recreational Water Quality,” MS Thesis, School of Energy, Environmental, Biological, and Medical Engineering, University of Cincinnati, Cincinnati, OH, p. 101
- Articles in Refereed Scientific Journals:
    - Motamarri, S. and D. L. Boccelli (2011) “Development of a Neural Based Biomarker Forecasting Tool to Classify Recreational Water Quality” in preparation for Water Research.

Report Follows

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

D. L. Boccelli (University of Cincinnati)

1. A progress report containing Problem and Research Objectives, Methodology, and Principal Findings and Significance

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. Additional studies will focus on exploring important explanatory values to develop the simplest model formulation that adequately represents the range of observed data.

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% minus the 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 preliminary results were followed by research that investigated the importance of the eighteen potential explanatory parameters for classifying the microbial safety of recreational surface waters through input selection techniques (backward stepwise regression for MLR; perturbation analysis for ANN; and variance gain method for LVQ). For the MLR approach, the important factors include the discharge in the river during the current and previous day, the amount of rainfall in the previous week, and the intensity of the storm event were observed to be important. For the ANN approach, the important explanatory variables were intensity of the storm event, average daily net radiation, time since storms greater than 0 and 1.0 inches, and the amount of rainfall in the previous day and week. For the LVQ approach, the important explanatory variables had to be separately investigated for the swimming and boating standards. For both standards, the intensity of storm event, average daily net radiation, and the time since storms greater than 0, 0.2, and 0.5 inches were considered to be important. For the swimming standard, the discharge from five days prior was also important. For the boating standard, the discharge of the current day was also important. For all three approaches the storm intensity was found to be important, and the discharge, rainfall amount, and time since certain rainfalls were found important in varying combinations. The net radiation appears to be important for ANN and LVQ, but not for the MLR.

From the standpoint of classification, throughout the input selection process, the ANN and LVQ approaches consistently outperformed the MLR algorithm. Additionally, the LVQ algorithm performed as good as or better than the ANN algorithm. These results suggest that the LVQ can be a useful algorithmic tool for quickly classifying recreational surface waters relative to the appropriate standard.

2. Publication citations

Publications

Motamarri, S. and D. L. Boccelli (2011) "Development of a Neural Based Biomarker Forecasting Tool to Classify Recreational Water Quality" in preparation for Water Research.

Motamarri, S. (2010) "Development of a Neural Based Biomarker Forecasting Tool to Classify Recreational Water Quality," MS Thesis, School of Energy, Environmental, Biological, and Medical Engineering, University of Cincinnati, Cincinnati, OH, p. 101

Motamarri, S. and D. L., Boccelli (2009). "The Development of a Neural-Based Biomarker Forecasting Tool for Classifying Recreational Water Quality" in Proceedings of the World Water and Environmental Resources Congress, ASCE, Kansas City, MO.

3. Number of students supported by the project (MS/PhD) as well as their majors

Srinivas Motamarri (MS in Environmental Engineering) – currently at the US EPA (Athens, GA)

4. Awards or Achievements

The current research performed through the Water Resources Center has resulted in a follow-up proposal submitted to the Metropolitan Sewer District (Cincinnati, OH) to continue this research. The proposal is

"Development of a Recreation Management Tool to Predict Microbial Water Quality for the Metropolitan Sewer District" (2010), PI: D. L. Boccelli, co-PI: S. G. Buchberger, University of Cincinnati ($153,073)