Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/watres

Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms

*Srinivas Motamarri, Dominic L. Boccelli**

School of Energy, Environmental, Biological and Medical Engineering, P.O. Box 210012, University of Cincinnati, Cincinnati, OH 45221-0012, USA

ARTICLE INFO

Article history:
Received 29 December 2011
Received in revised form
14 March 2012
Accepted 13 May 2012
Available online 23 May 2012

Keywords:
Surface water
Linear regression
Artificial neural network
Learning vector quantization
Classification
Fecal coliform

ABSTRACT

Users of recreational waters may be exposed to elevated pathogen levels through various point/non-point sources. Typical daily notifications rely on microbial analysis of indicator organisms (e.g., *Escherichia coli*) that require 18, or more, hours to provide an adequate response. Modeling approaches, such as multivariate linear regression (MLR) and artificial neural networks (ANN), have been utilized to provide quick predictions of microbial concentrations for classification purposes, but generally suffer from high false negative rates. This study introduces the use of learning vector quantization (LVQ) – a direct classification approach – for comparison with MLR and ANN approaches and integrates input selection for model development with respect to primary and secondary water quality standards within the Charles River Basin (Massachusetts, USA) using meteorologic, hydrologic, and microbial explanatory variables. Integrating input selection into model development showed that discharge variables were the most important explanatory variables while antecedent rainfall and time since previous events were also important. With respect to classification, all three models adequately represented the non-violated samples (>90%). The MLR approach had the highest false negative rates associated with classifying violated samples (41–62% vs 13–43% (ANN) and <16% (LVQ)) when using five or more explanatory variables. The ANN performance was more similar to LVQ when a larger number of explanatory variables were utilized, but the ANN performance degraded toward MLR performance as explanatory variables were removed. Overall, the use of LVQ as a direct classifier provided the best overall classification ability with respect to violated/non-violated samples for both standards.

1. Introduction

The use of recreational surface waters can pose a public health risk due to elevated pathogens resulting from the discharge of untreated, or partially treated, sewage and stormwater through various point and non-point sources. Within the U.S., the Clean Water Act provides the states with the authority to develop water quality standards to protect the nation's waterways and coastal regions. In 1986, the U.S. Environmental Protection Agency (USEPA) recommended the use of *Escherichia coli* and/or enterococci as fecal indicator organisms (FIO) to assess pathogen concentrations associated with gastrointestinal illnesses (USEPA, 1986). The Beaches Environmental Assessment and Coastal Health (BEACH) Act (2000) required the coastal and Great Lake states to adopt bacterial standards that provided the same, or better, levels of protection as the 1986

* *Corresponding author*. Tel.: +1 513 375 6901; fax: +1 513 556 4162.
E-mail addresses: srinivasiitr@gmail.com (S. Motamarri), dominic.boccelli@uc.edu (D.L. Boccelli).
0043-1354/$ – see front matter
doi:10.1016/j.watres.2012.05.023

recommendations for waters used for primary recreational activities. The criteria for *E. coli* and enterococci are specified for both a steady state geometric mean and four single sample maximum concentrations based on usage intensity (USEPA, 2004). While the geometric mean is useful for assessing the overall water quality, a single sample maximum concentration is best suited to inform recreational water closures.

Unfortunately, the typical approach for enumerating indicator bacteria to determine the single sample concentration is through sampling and analysis using USEPA approved methods, which require more than 18 h to develop a reportable result. During the time required to produce an analytical result, the public may be exposed to elevated pathogen levels. An expert panel recognized such a time delay as a limitation associated with using single sample measurements to inform recreational water closures, and supported the use of simple and/or mechanistic models for water quality prediction as part of a daily notification system (USEPA, 2007). For example, Maimone et al. (2007) used statistical data analysis to develop a real-time web-based system for classifying microbial water quality in the Schuylkill River (Pennsylvania, USA) that generates one of three classification levels based upon turbidity, flow, and rainfall data. The algorithm correctly classified 65% of the samples tested while generating higher classifications (a "false positive") in the other 35% of the samples. Other studies have focused on predicting bacteria concentrations – primarily using linear regression and artificial neural networks (ANNs) – for classifying the status of recreational waters. The following presents some of these studies focused on the prediction and classification associated with fresh water systems.

Several studies have utilized linear regression techniques to either explore the importance of various explanatory variables or to predict bacteria concentrations (Christensen et al., 2000; Desai et al., 2010; Ferguson et al., 1996; Hampson et al., 2010; Jagupilla et al., 2010; McCarthy et al., 2007; Schoonover and Lockaby, 2006). These studies have been used to represent fecal coliform or *E. coli* concentrations with explanatory variables ranging from land cover to population/livestock density to various combinations of hydrologic, meteorologic, and microbial variables. While these studies provide insight into the important factors associated with indicator bacteria concentrations, the predictive power of these models vary with reported R^2 values ranging from 0.17 to 0.99.

Additional studies have utilized linear regression to predict FIO concentrations using different combinations of hydrologic, meteorologic, and microbial variables and then classify the water quality based on the appropriate state water quality standard (Eleria and Vogel, 2005; Francy et al., 2006; Heberger et al., 2008; Hellweger, 2007). Table 1 summarizes the results of these studies including the FIO, number of explanatory variables, modeling approach, prediction and classification performance, and magnitude of water quality standard utilized for classification purposes. Eleria and Vogel (2005) (who also used logistic regression) and Francy et al. (2006) utilized the predicted FIO concentrations from linear regression models for direct classification. Heberger et al. (2008) estimated the probability that the predicted FIO concentration exceeded the water quality standard, and utilized a probability threshold to classify both violated and non-violated samples. Hellweger (2007) compared the prediction and classification performance of linear regression and a three-dimensional mechanistic hydrodynamic and water quality model, as well as ensemble models that included a 50–50 average of the regression and mechanistic model, and

Table 1 – Summary of previous research studies focused on utilizing linear regression and artificial neural networks (separated by a vertical space) for classifying surface water quality based on predicted fecal indicator organism (FIO) concentrations.

Study	FIO[a]	Explanatory variables	Modeling approach[b]	R^2	TN/TP[c] (%)	Standard (cfu/100 mL)
Eleria and Vogel (2005)	FC	23 different variables	LR	0.54–0.69	97/63	1000
			LogR	0.46–0.56	97/63	
Francy et al. (2006)	EC	9 different variables	LR	0.35–0.44	TN: 53–99	235
					TP: 26–93	
Heberger et al. (2008)	Ent	Precipitation;	LR	0.42–0.82	TN: 88, 84	61/305
		intra-event time; discharge			TP: 89, 100	
Hellweger (2007)	EC	Discharge; CSO;	LR	0.60	80/98	235
		wind speed/direction	Mechanistic		93/70	
			Ensemble (50/50)		97/77	
			Ensemble (max)		74/99	
Chandramouli et al. (2007)	FC	7 different variables	ANN	0.63–0.94	97/61	200
Mas and Ahlfeld (2007)	FC	7 different variables	LR		TP: 51/38	20/200
			LogR		TP: 58–75/46	
			ANN		TP: 61–81/46–62	
Tufail et al. (2008)	EC	Discharge; turbidity	LR	0.66–0.69		Three classes
			ANN	0.58–0.73	Overall 84–88	
			FFSGA	0.70		

a FIO – fecal indicator organism: FC – fecal coliform; EC – *E. coli*; Ent – Enterococci.
b LR – linear regression; LogR – logistic regression; ANN – artificial neural network; FFSGA – fixed functional set genetic algorithms.
c TN – true negative; TP – true positive.