Classification and regression tree (CART) analysis for indicator bacterial concentration prediction for a Californian coastal area

Hun-Kyun Bae, Betty H. Olson, Kuo-Lin Hsu and Soroosh Sorooshian

ABSTRACT

The study used existing indicator bacterial data and a number of physicochemical parameters that can be measured instantaneously to determine if a decision tree approach, especially classification and regression tree, could be used to predict bacterial concentrations in timely manner for beach closure management. Each indicator bacteria showed different tree structures and each had its own significant variables; Dissolved oxygen played an important role for both total coliform and fecal coliform and turbidity was the most important factor to predict concentrations of enterococci for decision tree approaches. Root mean squared error stayed between 5 and 6.5% of the average values of observations; RMSEs from each simulation, 0.25 for total coliform, 0.31 for fecal coliform, and 0.29 for enterococci. Estimations from tree structures would be regarded as a good representation of the actual data. In addition to results of the objective function, RMSE, 77.5% of actual value fell into the 95% of confidence interval of estimations for total coliform concentrations, 60% for fecal coliform concentrations, and 62.5% for enterococci concentrations. The approach showed reliable estimations for the majority of the data processed, although the method did not portray low concentrations of bacteria as well.

Key words | CART, decision tree, indicator bacteria, physical and chemical parameter

Hun-Kyun Bae (corresponding author)
Betty H. Olson
Kuo-Lin Hsu
Soroosh Sorooshian
Department of Civil and Environmental Engineering,
Henry Samueli School of Engineering,
University of California,
1368 Social Ecology II,
Irvine, CA 92697,
USA
E-mail: *baeh@uci.edu*;
bholson@uci.edu;
kuolinh@uci.edu;
soroosh@uci.edu

INTRODUCTION

Coastal areas are essential habitats for many species including those that are threatened and endangered. Population growth, however, along with growing urbanization in the much desirable coastal areas has resulted in increasing the amount of contaminants deposited in the ocean through the drainage system. From 1985–1994, 3,713 individuals became ill from 55 outbreaks involving recreational waters in the United States. Epidemiological studies showed a strong association between gastrointestinal illness and concentrations of groups of indicator bacteria, and an association also existed for eye, ear, nose and throat infections, but was less robust (Hunter 1997). To increase protection of public health, the state of California, passed the Oceans Protection Act and has entered into a cooperative effort with the other two western coastal states. This Act reinforces California's commitment to investing substantial resources and manpower to ensure both monitoring and maintenance of beach water quality. Coastal water quality agencies and managers rely upon indicator bacterial concentrations in the surf zone to protect beach-goers from exposure to waterborne disease.

Current monitoring systems cannot provide real-time monitoring results because of its testing time requirements, at least 18 hr to 24 hr Given that 70% of the bacteria in the water body are naturally cleaned after 24 hours, beach closure statements are issued after the fact (Christen 2002). The difference between the testing time and the natural purification process makes it difficult to identify the level of water quality on a real time basis and may affect the health of coastal communities and their economies because water

doi: 10.2166/wst.2010.842

bodies remain closed after the water has naturally returned to acceptable concentrations of indicator bacteria (SWRCB 2001). Therefore, inaccurate decisions on beach openings or postings/closures will greatly affect the public health and economy of coastal communities. Loss of local income and health costs associated with gastroenteritis in Los Angeles and Orange County beaches alone are estimated to be between 21 to 51 million dollars annually (Moriki & Karydis 1994; California 1998; Oftelie *et al.* 2000; SWRCB 2001; Surfrider Foundation 2003; USFW 2004). In addition to this timing gap, the methods are labor intensive and costly. The State of California manages more than 400 water quality monitoring sites at the beach areas and monitors these sites at least once a week. The annual cost of this monitoring system is approximately 1.3 million dollars and the total costs would reach approximately $2.3 million dollars if EPA's estimate on labor costs for water sample collection and analysis of $5,690.00 per beach site per year were included (NRDC 2005)

Effective methods which are cost and labor efficient and could provide indicator bacterial concentrations on a real time basis should be developed to manage the water quality. The primary objective of this study was to investigate a method which could predict bacterial concentration in real-time through the application of a Classification and Regression Tree (CART) analysis. Five different physical and chemical parameters identified in previous work on this site (Bae *et al.* 2009) were selected as input variables for CART to predict the three indicator bacteria used in the state of California, because of their association with bacteria and the ability to measure these variables quickly and easily. The present study is focused on the Aliso Creek Watershed, which is the main source area of pollution for Aliso Beach in Orange County, California.

MATERIALS AND METHODS

Site description

The study area, Aliso Creek Watershed, is located in the southern part of Orange County, California. Aliso Creek, its average discharge 6.7 ft^3/sec (cfs) (0.19 m^3/s), is directly connected to the Aliso Beach which means the Watershed drains directly onto the beach, greatly affecting the water quality of surf zone, where small children tend to recreate (Figure 1). This study focused on the high beach use periods of May through September. ArcView GIS version 3.3 and AGWA (Automated Geospatial Watershed Assessment) were used to delineate the study area.

Data

The Orange County Watershed and Coastal Resources Division (OCWCRD 2006) provided streamflow and water quality data. Water quality samples were taken only for dry seasons, May through September, from 2003 to 2005 at the sampling station ACJ01. Streamflow is recorded every five minutes at the gage 1,146 during the whole year from July 1999 to June 2006. The Ocean Water Protection Program operated by Orange County Health Care Agency (OCHCA 2006) provided bacterial concentration data from 1999 to 2005 for the Aliso Beach. Indicator bacteria samples for the beach area were taken at the outlet of Aliso Creek (station C1). The sampling intervals were irregular, but at least twice

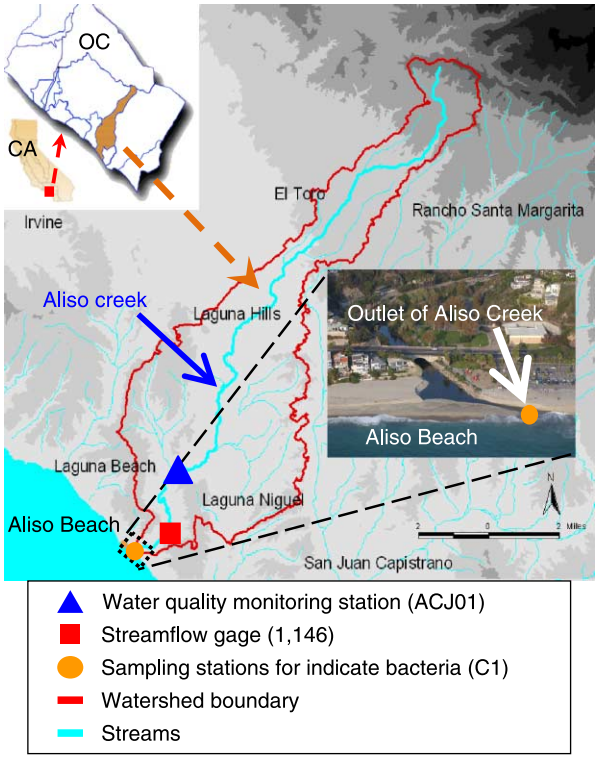

Figure 1 | The Study area, Aliso Creek Watershed, Orange County, California.

a week during the whole year. All samples were measured for concentrations of total coliform (TC), fecal coliform (FC) and enterococci (ENT).

Bacterial concentrations at C1, the outlet of Aliso Creek, showed 100 to 500 times lower than those at ACJ01 (data not shown). This fact confirms that bacterial load from upstream portions of Aliso Creek is the main source of indicator bacteria in the surf zone of the Aliso Beach. Origins of each bacterium for Aliso Creek watershed were defined from the previous, TC from natural systems such as soil, plants and FC and ENT from warm blooded animals, especially birds (Bae *et al.* 2009). Dilution effect, toxicity from salinity and precipitation of particle associated indicator bacteria might be the main reason for the differences between indicator bacterial concentrations at ACJ01 and those at C1 since flow rate of Aliso Creek is low. For reference, State standards of indicator bacteria for recreational water are 10,000 CFU/100 mL for TC, 400 CFU/100 mL for FC, 104 CFU/100 mL for ENT, and 1,000 CFU/100 mL for TC with the ratio of FC/TC $>$ 0.1 and the violation at the Aliso Beach for TC 0.13%, FC 29.52%, and ENT 70.17% (OCHCA 2006; OCWCRD 2006).

Decision tree analysis

A decision tree analysis is widely used for classification and prediction. A decision tree classifies data in the form of a tree structure which is generated from the use of training data in a top-down fashion or general-to-specific direction. The root node, initial state of a decision tree, is assigned all data. If data at the node of tree structure belong to the same class, so that no more decisions are needed, the node will be a leaf node which indicates the value of the target attribute (or class). If data at the node belong to two or more classes, such that the node has to be split, the node will be a decision node.

CART, one of decision tree methods, was adopted for this study. CART was developed by Breiman *et al.* (1984) and uses historical data to construct decision trees. The CART algorithm is a binary recursive tree structure which asks only the yes/no questions, so the parent nodes are always divided into two child nodes with searches for all possible variables and all possible values in order to find the best split. Then the process is repeated by treating each child node as a parent node until each node has maximum homogeneity such that further splitting is impossible or is limited by some criterion. CART develops a tree structure to split high-valued examples from low-valued examples by using least-squares difference from the sample mean (LS) or least absolute deviation from the sample median (LAD) to minimize an overall cost measure. Also, CART computes the importance of input variables using an ad-hoc ranking, scale of 0–100 based on the reduction of variance achieved. An ad-hoc ranking assigns different scales of units to the items, so more important items get the larger distance than less important ones. More important variables appear in the upper nodes and have significant impacts on the prediction of target values because movement down to one side at the upper nodes of the tree could bring much different results from those on the other side. On the other hand, less important variables, located at lower nodes of the tree, are less effective than those at upper nodes, but still related to outcomes of the tree analysis, so they are also considerable as factors. CART can easily handle both categorical and numerical variables; developing classification solution for categorical/discrete variables and a regression solution for numerical/continuous variables. It also has the ability to identify outliers as the splitting algorithm isolates outliers in an individual node (Timofeev 2004; CART 2006).

RESULTS AND DISCUSSION

Single variable of water quality vs. bacterial concentration

Figure 2 shows the relationship between physic-chemical parameters and indicator bacteria at the stations ACJ01. Truncated data were existed at 200,000 CFU/100 mL for summer of 2004 (data not shown), so samples for summer of 2005 seemed to be more diluted. Changes of streamflow were not significant, less than 1.0 cfs (0.028 m^3/sec). Streamflow seems to have relationships with bacterial concentrations, but many outliers remained and these outliers made it difficult to find a clear a relationship (Figure 2(a)–(c)). Figure 2(d)–(f) show the relationship between DO and bacterial concentration. Levels of DO in Aliso Creek cover a wide range from 2 mg/L to 14 mg/L and this is interesting

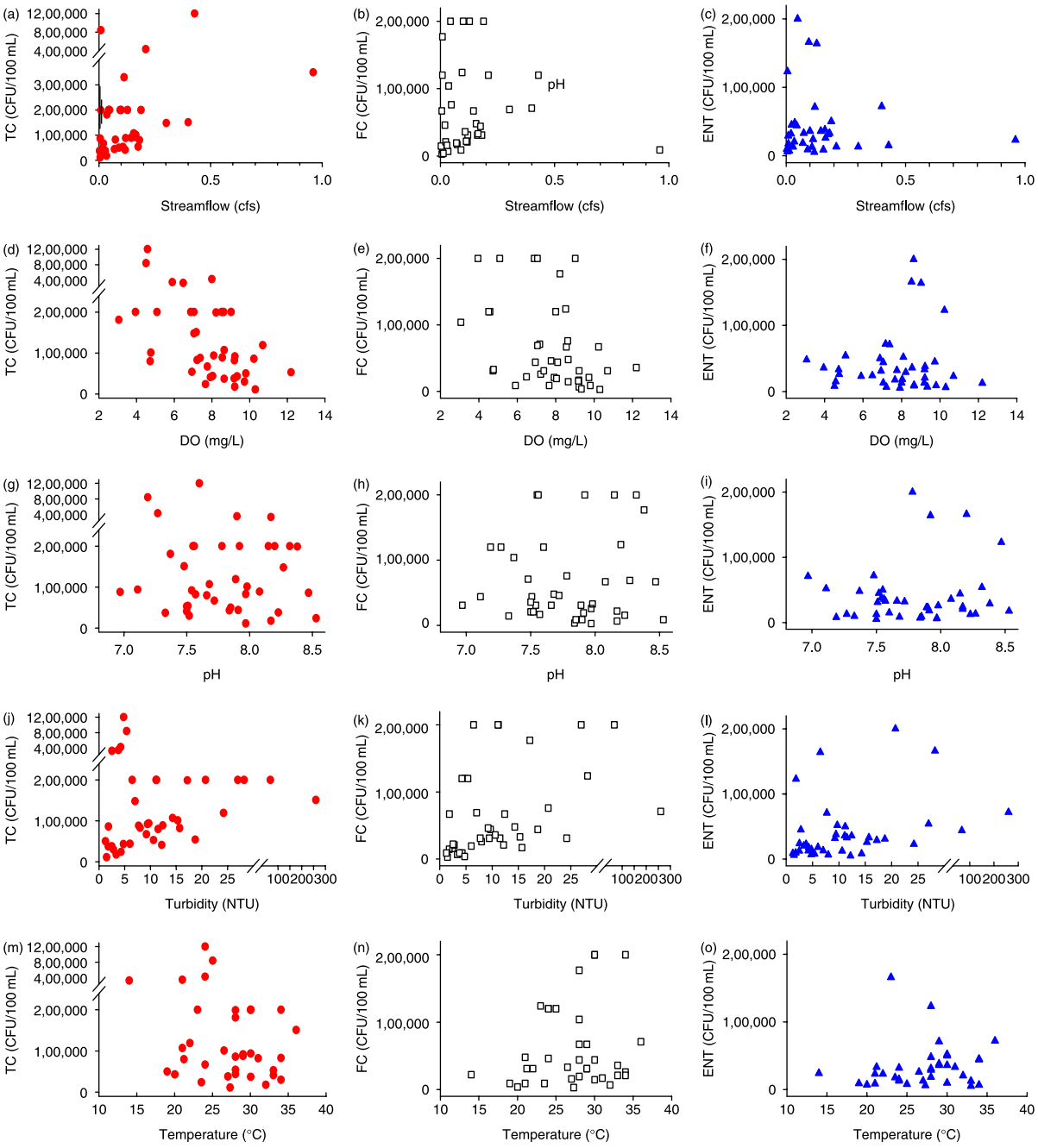

Figure 2 | Physico-chemical parameters vs. bacterial concentration at ACJ01.

if considering streamflow stays below 1 cfs (0.028 m^3/sec) with little variation. While indicator bacterial concentrations demonstrated relatively narrow ranges for samples taken within the same year, data compared across years showed markedly different concentration levels. DO levels were highly variable for samples collected within a year and also in across years samples. The level of DO roughly showed the negative patterns with indicator bacterial concentration, but outliers still existed and they made it difficult to find clear patterns. Truncated data where insufficient dilutions were removed to avoid prohibiting interpretation of the relationships. Nonetheless the relationship is negative

overall. Figure 2(g)–(i) show the relationship between pH and bacterial concentration. pH in Aliso Creek waters during summer seasons varied from neutral to alkaline (7.0 to 8.6). Ignoring truncated data and outliers indicated that pH showed an overall negative relationship with indicator bacterial concentrations. Relationships between turbidity and indicator bacterial concentrations were shown in Figure 2(j)–(l). Changes of turbidity generally tracked with bacterial concentrations since it is already reported that bacterial concentrations are associated with particles and sediments which parameters strongly related to turbidity (Auer & Niehaus 1993; Steets & Holden 2003). The study also showed that overall trends between turbidity and bacterial concentrations were positive relationships although the patterns were still affected by several outliers. It may be that turbidity and ENT demonstrated similar slopes, but at higher concentrations the intercept is shifted upward. Figure 2(m)–(o) show temperature and bacterial concentrations at ACJ01. Temperature showed complicated patterns with indicator bacteria. Temperature roughly had negative relationships with TC and ENT concentrations and positive relationship with FC, but relationships were unclear. Overall, clear patterns were not captured between physic-chemical parameters and indicator bacteria because of outliers.

Multi-variables of water quality vs. bacterial concentration

Multi-variables were considered together to find better relationships with bacterial concentrations since single variables didn't show clear relationships with bacterial concentration. Figure 3 shows results of one example, physicochemical parameters vs. TC concentrations sampled at ACJ01 after the datasets were re-organized with level of streamflow. Left side of the figures showed relationships between datasets matched to streamflow, lower than 1 cfs (0.028 m^3/sec) and right side those streamflow higher than 1 cfs (0.028 m^3/sec). Clear relationships were still not found because of outliers. However, partial relationships were shown since dividing datasets matched to the level of streamflow could separate outliers and bring better patterns between physicochemical parameters and indicator bacteria. Actually, datasets matched to streamflow <0.1 cfs (0.028 m^3/sec) had only one outlier and showed better relationships. In addition to getting better relationships, some variables showed both positive and negative patterns with indicator bacteria. Separating the dataset dependent on some variables could bring more detailed relationships, so the study adopted a decision tree analysis which could consider multiple variables.

Decision tree analysis

Decision tree analysis was used to activate all adopted physicochemical variables. CART Ex Version 6.0 was used for decision tree analysis (Salford Systems 2006). CART is designed to select two different modes, classification and regression, because each observation in the study has a different response and regression trees have response values while classification trees have classes.

Figure 4 showed tree structures for prediction of indicator bacteria concentrations from CART. Decision nodes including the root node (diamond) show the value of each physic-chemical parameter and leaf node (rectangular), the final level of bacterial concentration. For the decision tree analysis, more significant variables are located at the upper levels of the tree and less significant ones at the bottom of the tree. Therefore, DO played an important role for both TC and FC (Figure 4(a) and (b)) since DO located at the root node for trees of TC and FC predictions.

On the other hand, turbidity was the most important factor to predict concentrations of ENT (Figure 4(c)). CART didn't select pH for the tree of TC prediction (Figure 4(a)), temperature for FC (Figure 4(b)) and streamflow for ENT (Figure 4(c)), respectively. Therefore, those parameters were relatively less important to predict each indicator bacteria. The tree for TC prediction was smaller than trees for FC and ENT predictions; tree for TCprediction had four decision nodes, FC seven decision nodes and ENT six decision nodes. Overall, each indicator bacteria showed different tree structures and each had its own significant variables.

To validate each tree, the leave-one-out cross validation was adopted since each dataset was limited to separate into two different groups for training and validating tree structures (Shao 1993; DelSole & Shukla 2002). Figure 5 shows the process of leave-one-out cross validation. As shown in the figure, n numbers of the dataset were

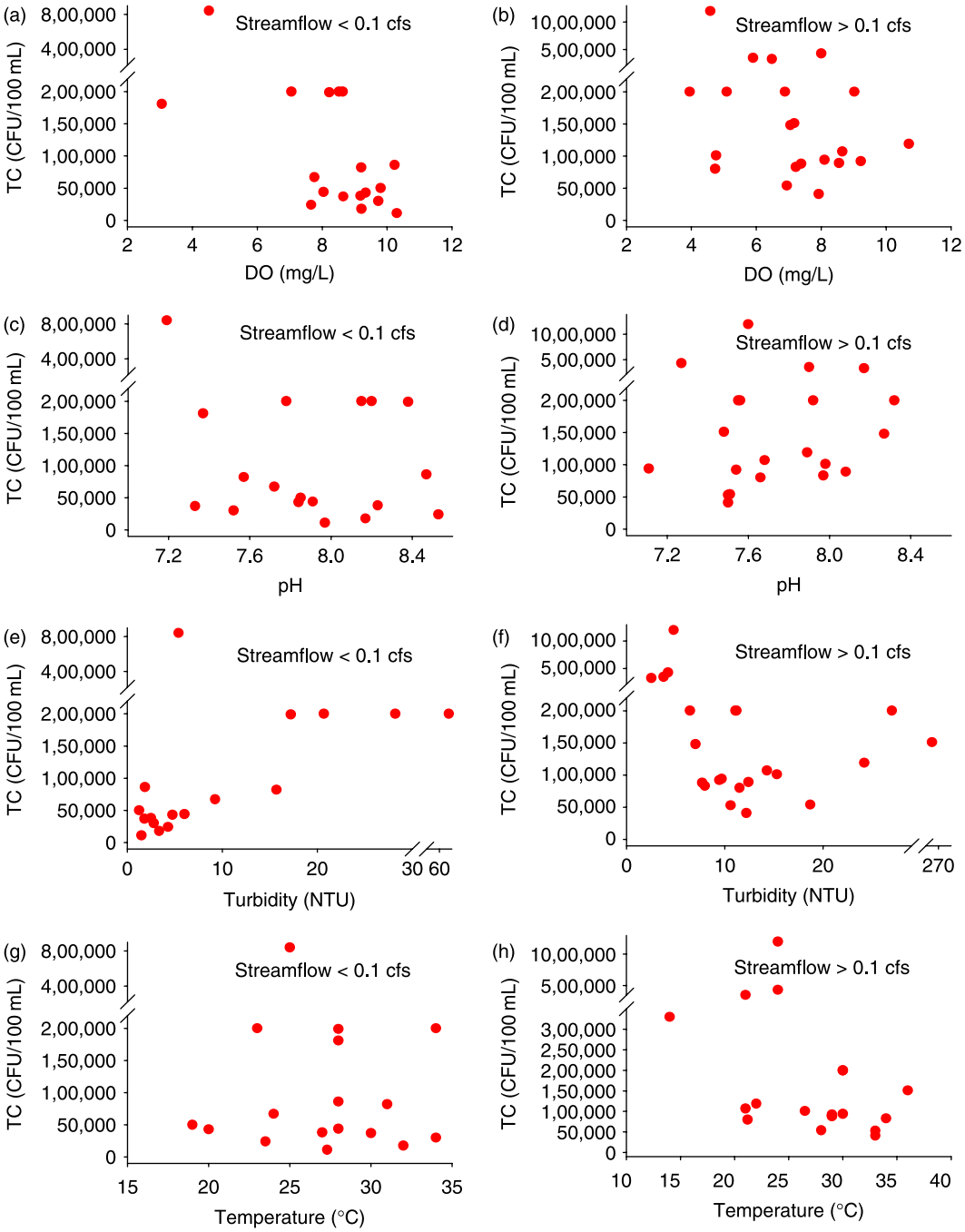

Figure 3 | Relationships between physic-chemical parameters and TC after re-organized data depending on the level of streamflow.

divided by two parts, $n - 1$ data points for training process and one data point for validation process, and the process was repeated until all data points were used for verification process. For example, the first data point was left for validation process and rest of other data, from the second data point to end of the data, were used for training process. And then, during next round, the second data point was separated for the validation process while rest of other data, the first data point + the third data point to end of the data, were used for training process. The process kept doing until

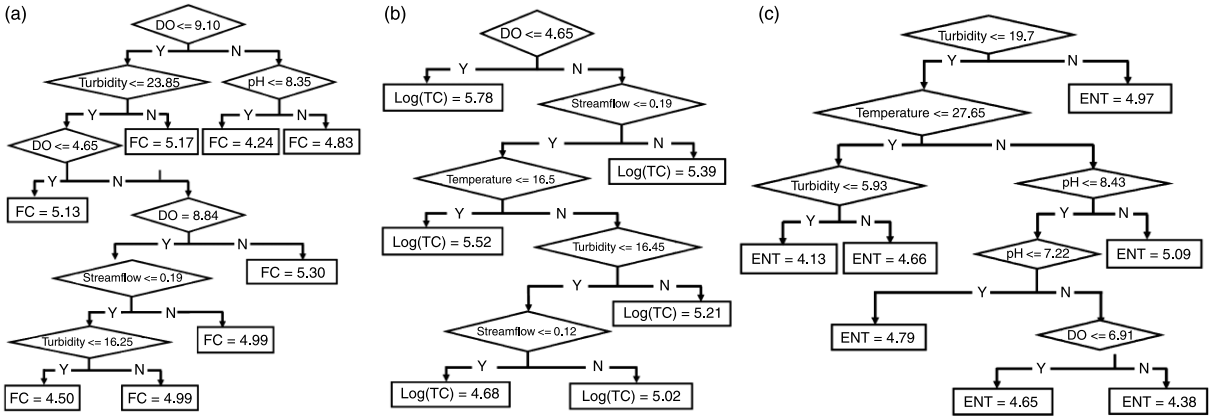

Figure 4 | Decision tree structures from CART for indicator bacterial concentration Prediction at ACJ01.

all data points were used for validation. The study, therefore, had $n - 1$ dataset for training processes and each trained tree was tested with leave-one-out data which was not included in training dataset. The root mean square error (RMSE) between estimations from each validation process and observations was calculated to measure the performance of tree structures for each indicator group.

Figure 6 showed leave-one-out cross validation results for decision tree approach for predicting bacterial concentrations at ACJ01. Bacterial concentration predictions from each tree structure showed reliable results. RMSE values stayed between 5 and 6.5% of the average value of observations; RMSEs from each simulation, 0.25 for TC, 0.31 for FC, and 0.29 for ENT. Estimations from tree structures would be regarded as a good representation of the actual data. In addition to results of objective function (RMSE), 77.5% of actual value fell into the 95% of confidence interval of estimations for TC concentrations, 60% for FC concentrations, and 62.5% for ENT concentrations. Therefore, estimations for TC concentrations showed the best fit to the actual event and estimations for ENT concentration were slightly better than those for FC concentration. Furthermore, all estimations showed several cases which did not capture extreme events (dashed ovals in Figure 6). The early stage of validation for TC concentration overestimated actual values, which were lower than other TC concentrations, while estimations for FC concentrations and ENT concentrations had missed several points of observations. Since CART is known for its ability to separate outliers, results are somewhat absurd. The probable reason for missing these points is that the number of training dataset was not good enough to cover those extreme events. Although each indicator bacterium had several extreme events, related input variables to those events were different, so validation of CART could not capture all cases since decision from CART structure depends on one variable at each step. In other words, CART approach of the study couldn't test cases related to all input variables since the dataset for the study had only 40 points. However, the approach still showed potential to predict bacterial concentrations as its performance was previously mentioned. Moreover, over- or underestimated

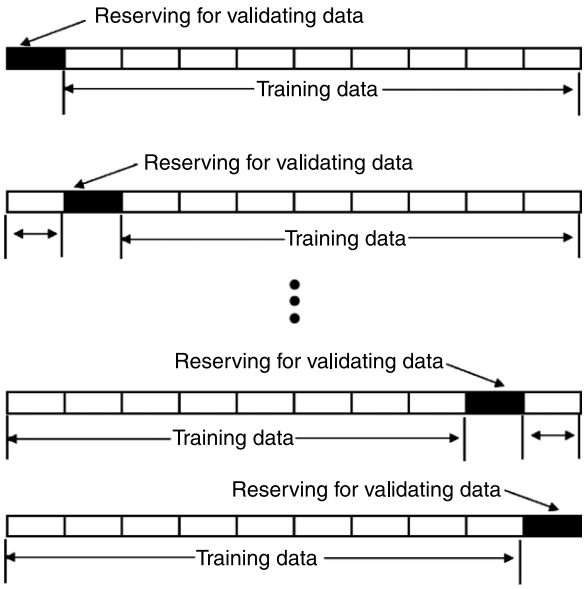

Figure 5 | Process of Leave-One-Out Cross Validation.

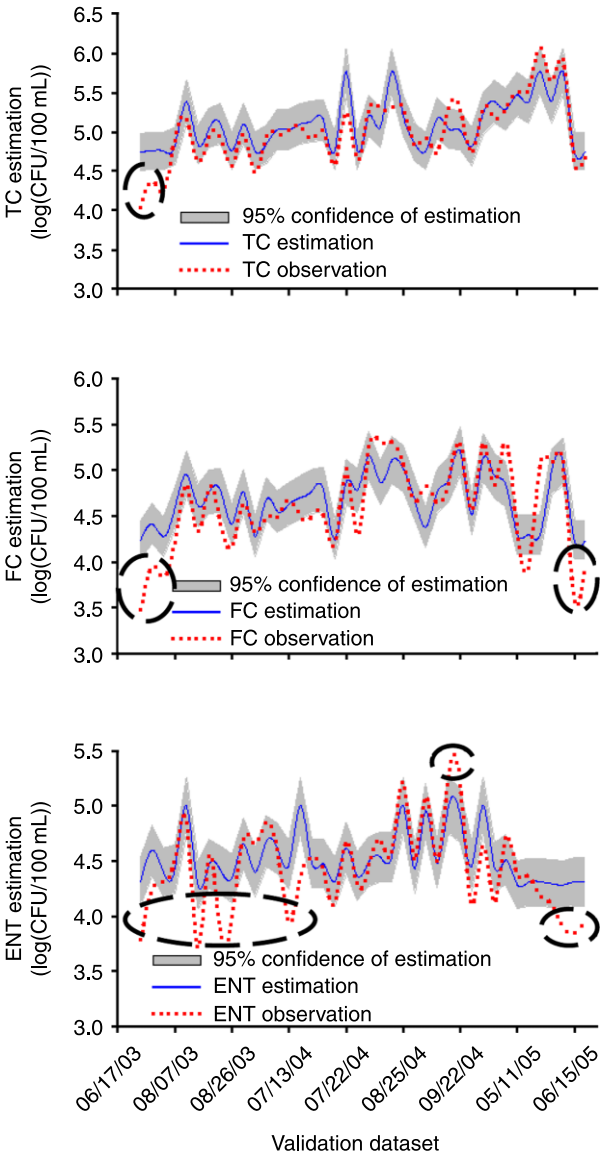

Figure 6 | Leave-one-out cross validation results for decision tree structures for indicator bacterial concentration prediction.

values were mostly for lower concentrations. Since fecal indicators standards are based on exceedences of established concentrations, the performance of tree structures for predicting indicator bacterial concentrations was reliable since they captured high peak events.

CONCLUSION

Relationships between bacterial concentration and water quality parameters are variable and can be influenced by several parameters simultaneously. This is shown by the partial relationships between single variables and indicator bacteria. Hence, variables may not be used alone for predicting concentrations. Selected variables in this study are considered to have relationships to bacterial growth as well as their survival. However, biological data in the environment could be affected by many different conditions which this study exemplified as single variables didn't show clear relationships with indicator bacteria concentrations. Once the datasets were divided by two groups dependent on one variable, other variables showed different patterns with indicator bacteria in each group, so that more relationships emerged after considering multiple variables. The study, therefore, adopted multiple variables with CART, one type of decision tree. Each indicator bacteria had its own significant water quality variables for decision trees and pH did not appear as a parameter to predict TC concentrations, temperature and streamflow did not appear in trees of FC and ENT predictions, respectively. However, all physicochemical parameters were required to build tree structures for all indicator bacteria. The cross validation was adopted to verify the performance of each tree structures and trees showed reliable results even though trees still had several points which were over- or under-estimated. Overall, the study showed the CART approach could provide a potential method to predict indicator bacterial concentrations using surrogate water quality variables which are easily and quickly measured. Information provided by CART simulations, therefore, might help decision makers to decide the conditions of water quality promptly. The analysis also suggested that different indicator bacteria had own significant input variables, so one should adopt input variables carefully if CART approach might be used for predicting indicator bacteria.

ACKNOWLEDGEMENTS

Partial financial support of this study is made available through research grants from National Science Foundation Sustainability of semi-Arid Hydrology and Riparian Areas (SAHRA: Grant Y414423) program.

REFERENCES

Auer, M. T. & Niehaus, S. L. 1993 Modeling fecal colifrom bacteria-I. Field and laboratory determination of loss kinetics. *Water Res.* **27**(4), 693–701.

Bae, H.-K., Olson, H. B., Hsu, K.-L. & Sorooshian, S. 2009 Identification and application of physical and chemical parameters to predict indicator bacterial concentration in a small Californian creek. *Water Environ. Res.* **81**(6), 633–640.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 Classification and Regression Trees (CART), Pacific Grove, Wadsworth, Monterrey, CA.

California State Assembly Bill: AB411 1998

CART (Classification and Regression Trees) 2006 http://www.statsoft.com/textbook/stcart.html

Christen, K. 2002 Making accurate water-quality determinations. *Environ. Sci. Technol.* **36**(19), 368A–369A.

DelSole, T. & Shukla, J. 2002 Linear prediction of Indian monsoon rainfall. *J. Clim.* **15**(24), 3645–3658.

Hunter, P. R. 1997 *Waterborne Disease*. John Wiley & Sons Ltd, Baffins Lane, Chichester, West Sussex, England.

Moriki, A. & Karydis, M. 1994 Application of multi-criteria choice-methods in assessing eutrophication. *Environ. Monit. Assess.* **33**(1), 1–18.

NRDC (Natural Resources Defense Council) 2005 http://www.nrdc.org/water/oceans/qttw.asp

OCHCA (Orange County Health Care Agency) webpage 2006 http://www.ocbeachinfo.com/downloads/index.htm

OCWCRD (Orange County Watershed & Coastal Resources Division) webpage 2006 http://www.ocwatersheds.com/watersheds/alisocreek_land_use.asp

Oftelie, S., Saltzstein, A., Gianos, P., Boyum, K., Rocke, R. & Mosallem, A. 2000 Infrastructure: latest survey finds orange county voters broadly similar to national survey respondents on the priority of cleaning up coastal waters, Technical Report, The Orange County Business Council.

Salford Systems 2006 CART Extended Edition Version 6.0, California Statistical Software, Inc., CA.

Shao, J. 1993 Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **88**(422), 486–494.

Steets, B. M. & Holden, P. A. 2003 A mechanistic model of runoff-associated fecal coliform fate and transport through a coastal logoon. *Water Res.* **37**(3), 589–608.

Surfrider Foundation 2003 http://www.surfrider.org/

SWRCB (State Water Resources Control Board, California Environmental Protection Agency) 2001 Source investigations of storm drain discharges causing exceedances of bacteriological standards.

Timofeev, R. 2004 Classification and Regression Trees (CART) Theory and Applications, Thesis, Center of Applied Statistics and Economics, Humboldt University, Berlin.

USFW (U.S. Fish & Wildlife Service) 2004 http://www.fws.gov/