Predictive Modeling of Microbial Indicators for Timely Beach Notifications and Advisories at Marine Beaches

Richard Zepp[1], Mike Cyterski[1], Marirosa Molina[1], Chris Fitzgerald[2], Gene Whelan[1], Rajbir Parmar[1], Kurt Wolfe[1], and Mike Galvin[1]

[1]US EPA, 960 College Station Rd., Athens GA 30605; [2]Student Services Authority

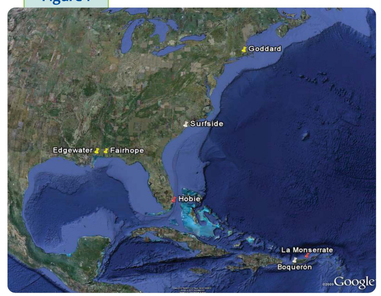
Introduction

Marine beaches are occasionally contaminated by unacceptably high levels of fecal indicator bacteria (FIB) that exceed EPA water quality criteria. Here we describe application of a recent version of the software package Virtual Beach tool (VB 3.0.6) to build and evaluate multiple linear regression (MLR) and GBM models to predict microbial water quality for selected marine beaches located in the eastern U.S. and Puerto Rico. Both culturable and qPCR methods were used to measure FIB (enterococci) concentrations at these beaches. Hydrometeorological and biogeochemical data also were obtained concurrently. Our objective was to compare results from statistical analyses of these data to compare marine and freshwater beaches in our date sets.

Methods

We used multiple linear regression analyses and the Generalized Boosted Regression Model (GBM) to examine data sets from seven marine beaches. The locations of the beaches are shown in Figure 1. These results were compared with analyses of data sets from five freshwater beaches all in the Great Lakes. At most sites, the response variable was either Enterococcus CFU or qPCR values. FIB data are primarily log-normally distributed, so we always $\log_{10}$ transformed the raw FIB measurements (CFU and qPCR) before model development. Site characteristics, the collection of water samples, laboratory bacterial measurement methods, and deployment of on-site instrumentation to collect environmental data are described several EPA reports and related journal articles. These are available on request.

Figure 1



Results & Discussion

MLR comparisons.

Twenty-seven independent variables (IVs) were found to be significant across analyses at the seven sites. Water temperature, humidity, and antecedent rainfall (typically cumulative over the past 48 hours) were most often significant. Salinity/conductivity, UV and/or other measurements of surface sunlight intensity, the number of birds seen on the beach, turbidity, and absorbance also appeared in many of the analyses. The qPCR model's adjusted R^2 exceeded that of the CFU model for only one of four data sets where nearly equal numbers of qPCR and CFU observations were taken. Comparisons of the MLR modeling results indicate that, on the basis of adjusted R^2 values for predicted versus observed levels of the FIB, model performance was better for the freshwater beaches than for the marine beaches (Figures 2-4). CFU results for adjusted R^2 are compared in Figure 2 and qPCR in Figure 3. In both cases it can be readily seen that the graph bars for the marine results are well below the averages for the Great Lakes beaches. As expected based on the adjusted R^2 results, the root mean square errors (RMSE) for the marine beaches were generally higher than for the freshwater beaches (Figure 4). The poorer performance of MLR models at marine beaches likely reflects the interplay of several factors such as the effects of currents and tides at the beaches (Grant and Sanders 2010), sunlight-induced inactivation (Boehm et al. 2009) and inputs of FIB from bird and dog droppings, bather shedding, runoff, groundwater and desorption from sand and decaying vegetation (Grant and Sanders 2010; Yamahara et al. 2007). Waves can reduce model performance; however, all but one of the marine beaches examined in the study were enclosed bays or estuaries that had subdued wave action. Boquerón Beach in Puerto Rico is a good example of an enclosed bay with low wave action. (Figure 5). On the other hand, Surfside Beach in South Carolina, with it's over 2 meter tidal changes and substantial wave action, provided one of the best model performances, especially with the qPCR results.

GBM comparisons.

The Generalized Boosted Regression Model (GBM, also known as a gradient boosting machine) uses binary decision rules (grouped together as a decision/regression "tree") to arrive at predictions of a response variable. For example, one such rule might be "If turbidity >= 15 NTU, increase/decrease the expected FIB concentration by some amount." The innovative aspect of GBM is that the algorithm doesn't solve for a single, complex decision tree: it builds a hierarchical set of simple trees, with each subsequent tree fit to the residual error from the previous tree. GBM avoids overfitting by developing each new tree based on a random set of these residual values.

While each tree is a simple structure, the long, linear combination of regression trees is more complicated. A negative aspect of a GBM model is that the model cannot easily be inspected graphically or expressed mathematically – it's something of a "black box." What it lacks in interpretability and transparency can often be made up in terms of prediction accuracy. Another noted aspect of GBM, unlike MLR and PLS, is that it handles non-linear relationships between the response and independent variables (IV's) without the need for transformations.

GBM is best used on datasets with > 50-100 observations; instability of the solution can occur on smaller datasets. The GBM method in Virtual Beach uses an algorithm for determining an optimal decision criterion for a fitted GBM model by striving for a balance between true negative and true positive outcomes. The GBM routine will not run if the dataset has no observations above the designated regulatory standard, so the user may be best served by defining the regulatory standard somewhere near the 75th percentile of the response variable distribution; this would provide a good number of observations on which to base the choice of a decision criterion.

GBM was used selectively to analyze data from several marine and freshwater beaches (Figure 6). Adjusted R^2 and RMSE are not options for evaluating performance of GBM models. Instead we used sensitivity, specificity and accuracy for our evaluations. Sensitivity is the number of correctly predicted exceedances over the total number of exceedances. Specificity gives the number of correctly predicted non-exceedances over the total number of non-exceedances. Accuracy gives the total number of correctly predicted values over the total count of the data set. Sensitivity results for three marine beaches, Boquerón Beach PR, Edgewater Beach MS, and Hobie Beach FL, are shown in Figure 6. For two of the beaches, sensitivity performance was significantly inferior to GBM results observed at freshwater beaches.

Figure 5

Figure 6

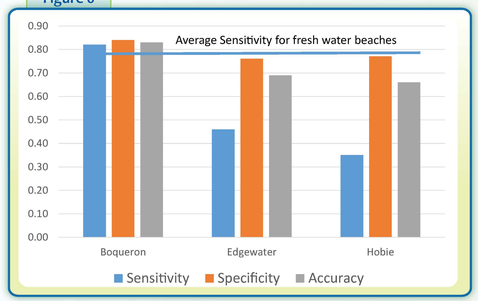


Figure 2

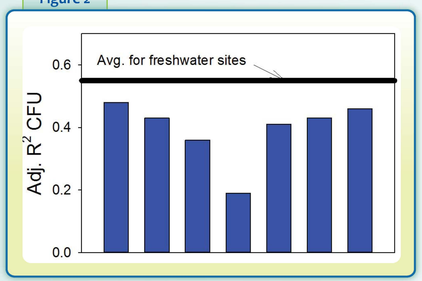


Figure 3

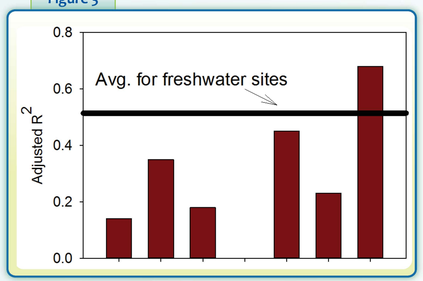
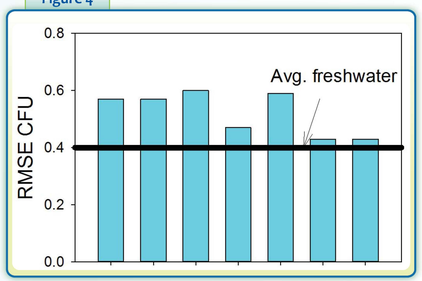


Figure 4



Conclusions

(1) Multiple linear regression models for marine beaches in the eastern United States and Puerto Rico performed more poorly than MLR models developed for selected Great Lakes beaches.

(2) The Generalized Boosted Regression Model (GBM) also indicated in a more limited way (based on sensitivity results) that model results for some marine beaches were inferior to Great Lakes beaches. GBM appears to be particularly more sensitive to difference in marine and freshwaters.

(3) These conclusions are based on limited data and analyses and additional comparisons of marine and freshwater beaches are needed to evaluate their generality.

References

Boehm, A.B., K.M. Yamahara, D.C. Love, B.M. Peterson, K. McNeill, and K.L. Nelson. 2009. Covariation and Photoinactivation of Traditional and Novel Indicator Organisms and Human Viruses at a Sewage-Impacted Marine Beach. Environmental Science & Technology 43(21):8046-8052.

Grant, S.B., and B. Sanders. 2010. The beach boundary layer: A framework for addressing recreational water quality impairment at enclosed beaches. Environmental Science & Technology in press.

Yamahara, K.M., B.A. Layton, A.E. Santoro, and A.B. Boehm. 2007. Beach sands along the California coast are diffuse sources of fecal bacteria to coastal waters. Environmental Science & Technology 41(13):4515-4521.