Comparing methods for predicting health advisories for beach water

Wesley Brooks, Rebecca Carvin, Steve Corsi

Abstract

Pithy, concise and informative. May bring the reader to tears due to the beauty of it.

Introduction

With input from the US Environmental Protection Agency, the state of Wisconsin has established regulatory standards for beach water quality, which states that a warning is to be posted when the concentration of E. coli exceeds 235 CFU / 100 mL. (Is that statement correct?) The goal of modeling the bacterial concentration is to predict in advance when the concentration will exceed the limit.

Methods

The availability of large data sets for building regression models to predict the bacterial counts in beach water is both an opportunity and a challenge.

Data Sources

Possibly move this to the end of the section

Which sites

Where are they

What specific sources sources of data (plug EnDDAT)

Will include a map and tables

Definitions

At any site, denote the predictor variables by X, which is an $n \times p$ matrix where n is the number of observations and p is the number of predictors. The vector of n observations of bacterial concentration is denoted y. The mathematical model relating y to X is the function $\mu(X, y)$. Denote the regulatory standard by δ and the decision threshold by $\hat{\delta}$.

Listing of specific statistical techniques

Fourteen different regression modeling techniques were considered. Each technique uses one of five modeling algorithms: GBM, the adaptive lasso, the genetic algorithm, PLS, or sparse PLS. Each technique is applied to either continuous or binary regression and to either variable selection and model estimation, or variable selection only.

Continuous vs. binary regression The goal of predicting exceednaces of the water quality standard is approached in two ways: one is to predict the bacterial concentration and then compare the prediction to a threshold, which is referred to as continuous modeling. The other is referred to as binary modeling, in which we predict the state of the binary indicator z_i:

$$z_i = I(y_i > \delta)$$

The indicator is coded as zero when the concetration is below the regulatory standard and one when the concentration exceeds the standard. All of the binary modeling techniques herein use logistic regression, which uses the logistic link function g to translate $p_i = E(z_i)$ - the probability that the ith observation is an exceedance - into an unbounded quantity.

$$g(p_i) = \log \frac{p_i}{1 - p_i}$$

Weighting of observations in binary regression A weighting scheme was implemented for some of the binary regression techniques. In the weighting scheme, observations were given weights w_i where:

$$w_i = (y_i - \delta)/\hat{sd}(y)\hat{sd}(y) = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n\bar{y}} = \sum_{i=1}^{n} y_i/n$$

That is, the weights are equal to the number of standard deviations that the observed concentration lies from the regulatory threshold δ. Any technique that was implemented with this weighting scheme was separately implemented without any weighting of the observations. The techniques are thus labeled weighted and unweighted, respectively.

Modeling or selection only The regression techniques adaptive lasso and sparse PLS include a variable selection step. , Some methods are labeled "select", which means that they are used for variable selection only. In these cases, once the predictor variables are selected, the regression model using those predictors is estimated using ordinary least squares for the continuous regression techniques, or ordinary logistic regression for the binary regression techniques.

GBM

GBM refers to the gradient boosting machine (GBM) (Friedman 2001). A GBM model is a so-called random forest model - a collection of many regression trees. Prediction is done by averaging the outputs of the trees. Two GBM-based techniques are explored - we refer to them as GBM and GBMCV. The difference is in how the optimal number of trees is determined - GBMCV selects the number of trees in a model using leave-one-out CV, while GBM uses the so-called out-of-bag (OOB) error estimate. The CV method is much slower (it has to construct as many random forests as there are observations, while the OOB method only requires computing a single random forest) but GBMCV should more accurately estimate the prediction error. All the GBM and GBMCV models share the following settings:

Number of trees: 10000

Shrinkage parameter: 0.0005

Minimum observations per node: 5

Depth of each tree: 5

Bagging fraction: 0.5

Adaptive Lasso

The adaptive lasso (Zou 2006) is a regression method that simultaneously selects relevant predictors and estimates their coefficients by adding a penalty to the sum of the squared residuals. For continuous modeling techniques the adaptive lasso selects the predictors for linear regression, estimating $\hat{\boldsymbol{\beta}}$ minimize the criterion

$$\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}\frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma},$$

where λ is a tuning parameter and $\tilde{\boldsymbol{\beta}}$ is a consistent estimate of the regression coefficients. For binary modeling, the adaptive lasso maximizes the penalized log-likelihood

$$\sum_{i=1}^{n}\left[-(1-y_i)X_i\beta - \log\left\{1+\exp\left(-X_i\beta\right)\right\}\right] + \lambda\sum_{j=1}^{p}\frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma},$$

where, as was the case for continuous response, $\tilde{\boldsymbol{\beta}}$ is a consistent estimate of the regression coefficients, which is calculated by ordinary logistic regression.

In this work, $\gamma = 1$, $\tilde{\boldsymbol{\beta}}$ are estimated individually by a univariate linear or logistic regression (it is necessary to estimate the coefficients individually because there are usually more covariates than observations), and the adaptive lasso tuning parameter λ is selected to minimize the AICc (Hurvich, Simonoff, and Tsai 1998).

Five of the modeling techniques were based on the adaptive lasso - one for continuous response (AL), and four for binary response (AL-logistic-weighted, AL-logistic-unweighted, AL-logistic-weighted-select, AL-logistic-unweighted-select). The four binary response techniques are the combination of weighted versus unweighted, and selection-only versus selection-and-estimation.

Genetic algorithm

The genetic algorithm (Fogel 1998) is a variable-selection method that works by analogy to natural selection, where so-called chromosomes represent regression models. A variable is included in the model if the corresponding element of the chromosome is one, but not otherwise. Chromosomes are produced in successive generations, where the first generation is produced randomly and subsequent generations are produced by combining chromosomes from the current generation, with additional random drift. The chance that a chromosome in the current generation will produce offspring in the next generation is an increasing function of its fitness. The fitness of each chromosome is calculated by the corrected Akaike Information Criterion (AICc) (Akaike 1973; Hurvich and Tsai 1989).

The implementations in this study used 100 generations, with each generation consisting of 200 chromosomes. The genetic algorithm method (GA) is the default for linear regression modeling in Virtual Beach (Cyterski et al. 2013). This study also investigates two genetic algorithm methods for logistic regression: one weighted (GA-logistic-weighted) and one unweighted (GA-logistic-unweighted).

PLS

Partial least squares (PLS) regression is a tool for building regression models with many covariates (Wold, Sjostrom, and Eriksson 2001). PLS works by decomposing the covariates into mutually orthogonal components, with the components then used as the variables in a regression model. This is similar to principal components regression (PCR), but the way PLS components are chosen ensures that they are aligned with the model output. On the other hand, PCR is sometimes criticised for decomposing the covariates into components that are unrelated to the model's output. To use PLS, one must decide how many components to use in the model. This study follows the method described in (W. R. Brooks, Fienen, and Corsi 2013), using the PRESS statistic to select the number of components.

SPLS

Sparse PLS (SPLS) combines the orthogonal decompositions of PLS with the sparsity of lasso-type variable selection (Chun and Keles 2007). To do so, SPLS uses two tuning parameters: one that controls the number of orthogonal components and one that controls the lasso-type penalty. The optimal parameters are those that minimize the mean squared prediction error (MSEP) over a two-dimensional grid search. The MSEP is estimated by 10-fold cross-validation. SPLS was used for both selection-and-estimation (SPLS) and selection-only (SPLS-select).

Implementation for beach regression

The response variable for our continuous regression models is the base-10 logarithm of the E. coli concentration. For the binary regression models, the response variable is an indicator of whether the concentration exceeds the regulatory threshold $\delta = 235$ CFU/mL. Transformations were applied to some of the data during pre-processing: the beach water turbidity and the discharge of tributaries near each beach were log-transformed, and rainfall variables were all square root transformed.

Include a table with pre/post processing discussion

Cross Validation

Our assessment of the modeling techniques is based on their performance in predicting exceedances of the regulatory standard. Two types of cross validation was used to measure the performance in prediction: leave-one-out (LOO) and leave-one-year-out (LOYO). In LOO CV, one observation is held out for validation while the rest of the data is used to train a model. The model is used to predict the concentration of that held out observation, and the process is repeated for each observation. Each cycle of LOYO CV holds out an entire year's worth of data for validation instead of a single observation. It is intended to approximate the performance of the modeling technique under a typical use case: a new model is estimated before the start of each annual beach season and then used for predicting exceedances during the season. The LOYO models in this study were estimated using all the available data except for the held out year, even that from future years. So for instance the 2012 models were estimated using the 2010-2011 and 2013 data.

Some methods also used cross-validation internally to select tuning parameters. In those cases the internal CV was done using only the model data, and never looking at the held-out observation(s). This process is separate from - and does not affect - the CV to assess predictive performance.

Comparing methods, and quantifying uncertainty in the ranks

Results were compiled into one table for each site, such as the one below which cotains the results of running the contest at Hika. Each observation corresponds to a row in the table. The results table has a column for the observed log *E. coli* concentration and, for each method, columns for the threshold and predicted concentration by LOO CV and for the threshold and predicted concentration by LOYO CV.

Row	Actual	`gbm` (LOO)	`gbm` (LOYO)	...	`adapt` (LOO)	`adapt` (LOYO)
1	2.54	2.35	2.22	...	2.29	2.55
2	2.59	1.87	1.79	...	1.91	1.23
3	2.73	1.97	2.22	...	1.79	2.09
⋮	⋮	⋮	⋮	...	⋮	⋮
166	1.57	1.93	2.06	...	1.83	2.07

Row	Actual	gbm (LOO)	gbm (LOYO)	...	adapt (LOO)	adapt (LOYO)
167	3.38	1.84	2.01	...	1.80	1.71

Our goal is to compare the different modeling techniques. We make comparisons on the basis of rank because the modeling metrics may not be comparable between sites (because, for instance, different sites have different numbers of observations). If it is possible to compute a metric of model performance, then it is possible to rank the techniques based on that metric. Our approach is to rank the techniques at each site (higher is better), and then average the ranks of each method.

In order tocompare techniques, we must quantify the uncertainty in the rankings, which we do via the bootstrap. Since the AUROC and PRESS rankings are functions of the result tables, the bootstrap procedure is carried out by resampling the rows of each results table, and the AUROC and ROC rankings were recalculated for each bootstrap sample. We used 1001 bootstrap samples of each results table in the analysis that follows.

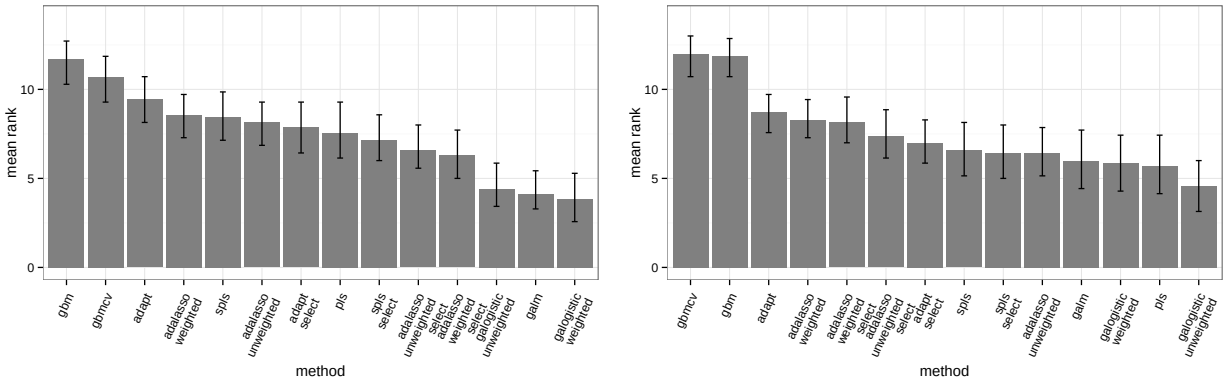
Results

AUROC

The receiver operating characteristic (ROC) curve is a graphical display of how well predictions are sorted. Each point on the curve represents the model's performance for a specific choice of the threshold $\hat{\delta}$ in terms of specificity and sensitivity. A model with well-sorted predictions is one where the predicted bacterial contentration (or the predicted probability of exeedance in the case of a binary regression model) is consistently greater when the true concentration exceeds the regulatory standard than when it does not. In such a case, the model may have both good sensitivity and good specificity over a wide range of thresholds $\hat{\delta}$.

The area under the ROC curve (AUROC) summarizes the model's performance over the range of possible thresholds. A model which perfectly separates exceedances from non-exceedances in prediction has an AUROC of one, while a model that predicts exceedances no better than a coin flip has an AUROC of 0.5.

The methods are ranked at each site by AUROC and a mean rank (across sites) is computed for each method. The mean LOO and LOYO ranks are plotted in Figure [fig:auroc-boxplot]. The top three techniques were GBM, GBM-CV and adaptive lasso. In order to facilitate a pairwise comparison between modeling methods, Tables [table:auroc.pairs.annual] (for the leave-one-year-out analysis) and [table:auroc.pairs] (for the leave-one-out analysis) show the frequency that the mean AUROC rank of GBM, GBM-CV, or the adaptive lasso exceeded each of the other modeling methods.



% latex table generated in R 3.1.1 by xtable 1.7-3 package % Wed Jul 30 19:32:17 2014

	gbmcv	adapt	adalasso weighted	spls	adalasso unweighted	adapt select	pls	spls select	adalasso unweighted select	adalasso weighted select	galogistic unweighted	galm	galogistic weighted
gbm	0.94	0.97	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
gbmcv		0.83	0.96	0.95	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00
adapt			0.76	0.80	0.84	0.98	0.91	0.97	0.98	0.98	1.00	1.00	1.00

Table 2: Under leave-one-year-out cross validation, how often the mean AUROC rank of GBM, GBMCV, or the adaptive lasso (in the rows) exceeded that of the other methods (in the columns).

% latex table generated in R 3.1.1 by xtable 1.7-3 package % Wed Jul 30 19:32:17 2014

	gbm	adapt	adalasso weighted	adalasso weighted select	adalasso unweighted select	adapt select	spls	spls select	adalasso unweighted	galm	galogistic weighted	pls	galogistic unweighted
gbmcv	0.52	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
gbm		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
adapt			0.61	0.64	0.82	1.00	0.96	0.98	0.96	0.98	0.99	0.98	1.00

Table 3: Under leave-one-out cross validation, how often the mean AUROC rank of GBM, GBMCV, or the adaptive lasso (in the rows) exceeded that of the other methods (in the columns).

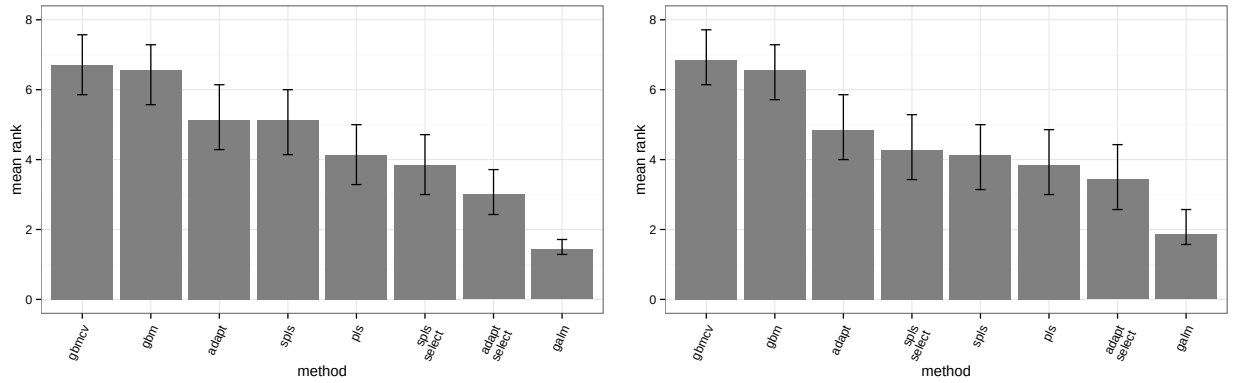
PRESS

While AUROC is concerned with how well the model's predictions separate exceedances from non-exceedances, the predictive error sum of squares (PRESS) measures how well a model's predictions match the observed bacterial concentration. The PRESS can only be computed for continuous regression methods. Let the model's predictions be denoted $\tilde{y}_i$ and letting the actual observed bacterial concentrations be denoted y_i for $i = 1, \ldots, n$ where n is the total number of predictions. Then PRESS computed as follows:

$$\text{PRESS} = \sum_{i=1}^{n} \left(\tilde{y}_i - y_i\right)^2 .$$

The PRESS statistic is of interest because a good model should accurately predict the bacterial concentration, and a model that more accurately predicts the concentration may be easier to trust. However, the AUROC is a more important as a metric of model performance than the PRESS because it directly measures the ability of a model to separate exceedances from non-exceedances. That said, we expect the two statistics to usually agree on which modeling methods are the best.

The rankings of the methods by PRESS are plotted in Figure [fig:press-boxplot]. The top three techniques under both LOO and LOYO analysis were GBM, GBM-CV and the adaptive lasso. In order to facilitate a pairwise comparison between modeling methods, Tables [table:auroc.pairs.annual] (for the leave-one-year-out analysis) and [table:auroc.pairs] (for the leave-one-out analysis) show the frequency that the mean AUROC rank of GBM, GBM-CV, or the adaptive lasso exceeded each of the other modeling methods.



% latex table generated in R 3.1.1 by xtable 1.7-3 package % Wed Jul 30 19:32:17 2014

	gbm	adapt	spls	pls	spls select	adapt select	galm
gbmcv	0.58	0.96	0.97	1.00	1.00	1.00	1.00
gbm		0.94	0.95	1.00	1.00	1.00	1.00
adapt			0.51	0.89	0.93	1.00	1.00

Table 4: Under leave-one-year-out cross validation, how often the mean PRESS rank of GBM, GBMCV, or the adaptive lasso (in the rows) exceeded that of the other methods (in the columns).

% latex table generated in R 3.1.1 by xtable 1.7-3 package % Wed Jul 30 19:32:17 2014

	gbm	adapt	spls select	spls	pls	adapt select	galm
gbmcv	0.65	1.00	1.00	1.00	1.00	1.00	1.00
gbm		0.97	0.99	1.00	1.00	1.00	1.00
adapt			0.71	0.80	0.85	0.94	1.00

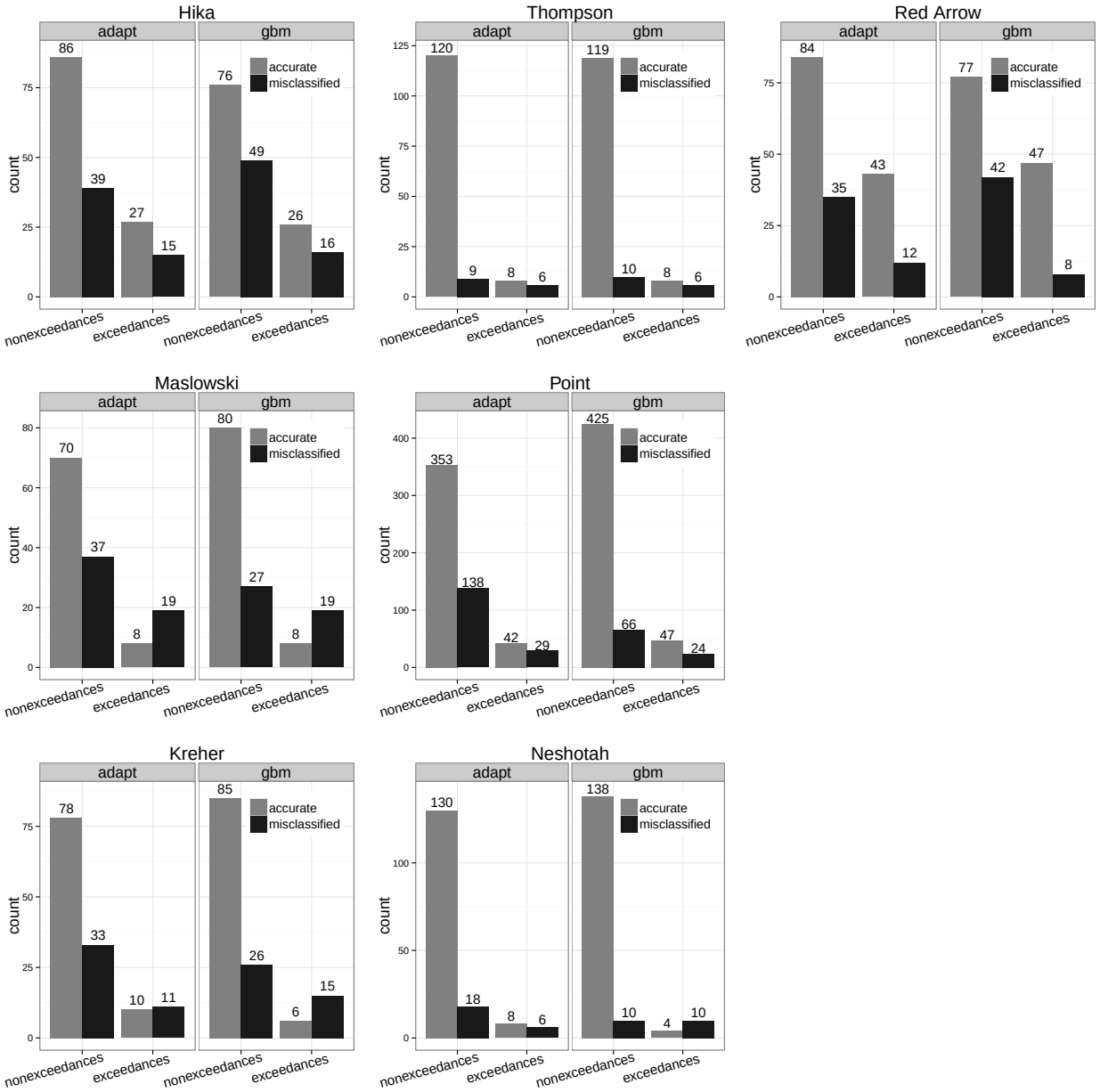
Table 5: Under leave-one-out cross validation, how often the mean PRESS rank of GBM, GBMCV, or the adaptive lasso (in the rows) exceeded that of the other methods (in the columns).

Classification of responses

While the AUROC is an important metric that should guide the choice of which modeling method to use, it measures the average performance over the possible range of thresholds. The real-world performance of any model for predicting exceedances will be measured by how well it distinguishes between exceedances and nonexceedances at the one specific threshold. Because LOYO CV was used to simulate real-world application of the modeling methods, it seems natural to evaluate the models based on the accuracy of their decisions when provided with a realistic decision threshold.

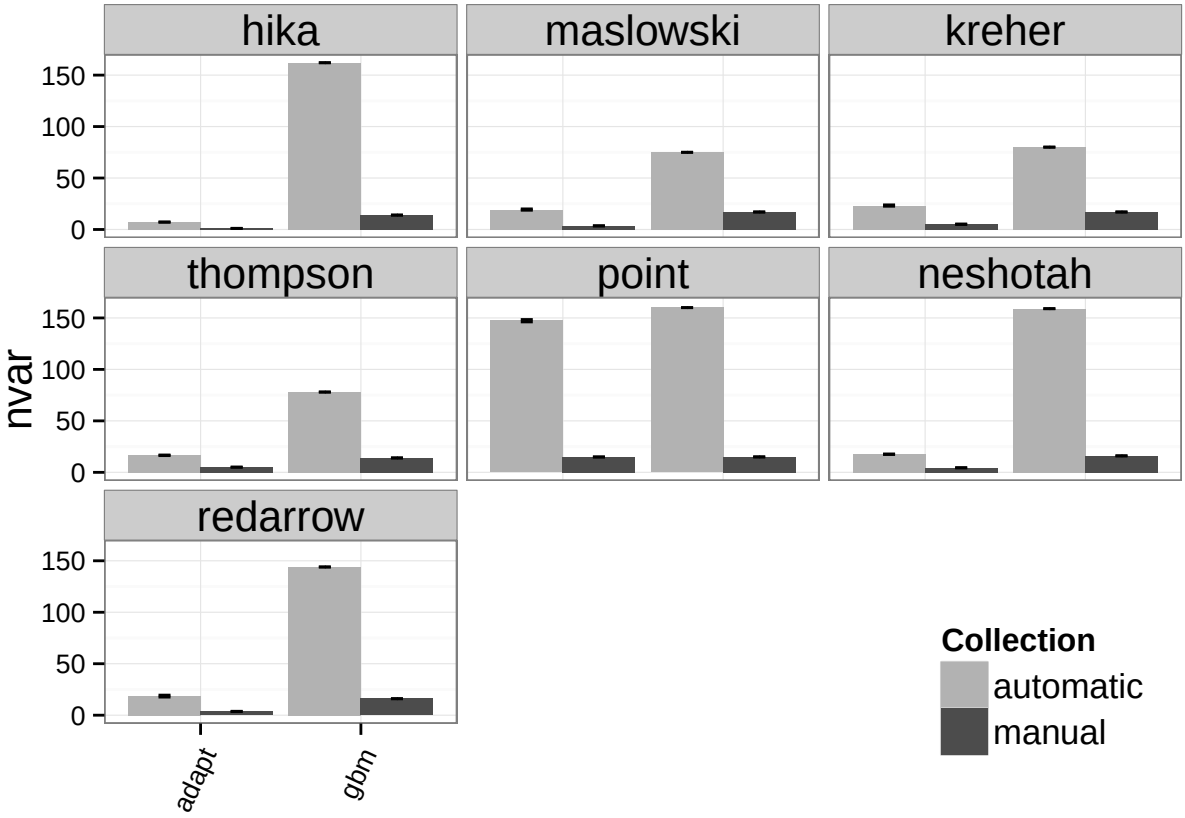
Intuitively, the decision threshold should adapt to the conditions that are observed in the beach's training data. If, for instance, exceedances are rare in the training data, then we expect few exceedances in the future, and should set the threshold high to reflect this prior expectation. On the other hand, if the bacterial concentration often exceeds the regulatory standard, then the decision threshold should be set lower in order to properly flag more of those exceedances. This intuition was encoded into how the decision threshold was set for the LOYO models. Specifically, the decision threshold was set to the q th quantile of the fitted values of non-exceedances in the training set, where 1-q is the proportion of exceednaces in the training set.

Since GBM, GBM-CV, and adaptive lasso were the top three techniques by both PRESS and AUROC, it seems like the best method to use will be one of these. The mean rankings average the results across all seven sites, and neither PRESS nor AUROC reflect the performance of a model under a particular choice of the decision criterion, so the results we've seen so far don't convey how often a model would indicate a correct decision if it were used to decide whether an advisory should or should not be posted at the beaches. In Figure [fig:counts-barcharts], we look at the counts on a per-beach basis of four categories of decisions: true positives (correctly posting an advisory), true negatives (correctly not posting an advisory), false positives (wrongly posting an advisory) and false negatives (wrongly not posting an advisory). The results are for GBM and adaptive lasso because the GBM and GBM-CV methods are so similar. In most cases, the counts are similar between the two techniques, with GBM tending to make a few more correct decisions. There are exceptions where adaptive lasso makes more correct decisions (e.g., Hika and Red Arrow).



Variable selection

Several of the methods we tested do variable selection to pare the number of covariates. We look here at how many variables were used in the adaptive lasso models compared to the GBM models, which use all the available covariates. The variable counts are displayed in Figure [fig:varselect-barchart]. At most of the sites, the adaptive lasso uses far fewer covariates than GBM, but at Point the adaptive lasso uses almost all of the available covariates. That's because the selection criterion we used (corrected AIC) is intended to minimize prediction error. As the amount of data increases (Point has far more observations than the other sites), we accumulate enough information to begin to discern the effect even of covariates that are only slightly correlated with the response. As our dataset grows, then, we should expect more covariates to be selected for the model.



Discussion

In general, the GBM, and GBMCV, and AL techniques produced comparable results that were superior to the other techniques in terms of predictive performance. Since the GBMCV models take much longer to compute than the others, we will not include them in our more detailed analysis of the modeling results.

Which type of model is generally the best?

Under what conditions do some outperform others?

Relative value of overall best model versus methods that help trim variables? e.g. how valuable is it to reduce number of predictors? Further, which variables get cut? Expensive ones? Cheap ones?

How important is computational expense? Only an issue for model fitting — not prediction, but worth quantifying. E.g. if GBM with cross validation takes hours, how much better?

Model tuning for GBM versus GBM-CV –> notes on how GBM is faster with similar performance (e.g. CV is overkill maybe)

Acknowledgments

References

Akaike, Hirotugu. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory*, 267–81. Akademinai Kiado.

Brooks, Wesley R., Michael N. Fienen, and Steven R. Corsi. 2013. "Partial Least Squares for Efficient Models of Fecal Indicator Bacteria on Great Lakes Beaches." *Journal of Environmental Mana* 114: 470–75.

Chun, Hyonho, and Sunduz Keles. 2007. "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection."

Cyterski, Mike, Wesley Brooks, Mike Galvin, Kirt Wolfe, Rebecca Carvin, Tonia Roddick, Michael N. Fienen, and Steven R Corsi. 2013. *Virtual Beach 3: user's Guide*. United States Environmental Protection Agency.

Fogel, David B. 1998. *Evolutionary Computation: the Fossil Record*. Wiley-IEEE Press.

Friedman, Jerome. 2001. "Greedy Function Approximation: a Gradient Boosting Machine." *The Annals of Statistics* 29: 1189–1232.

Hurvich, Clifford M., and Chih-Ling Tsai. 1989. "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76: 297–307.

Hurvich, Clifford M., Jeffrey S. Simonoff, and Chih-Ling Tsai. 1998. "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." *Journal of the Royal Statistical Society: Series B (Methodology)* 60: 271–93.

Wold, Svante, Michael Sjostrom, and Lennart Eriksson. 2001. "PLS-Regression: a Basic Tool of Chemometrics." *Chemometrics and Intelligent Laboratory Systems* 58: 109–30.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101: 1418–29.