

# Comparing methods for predicting health advisories for beach water

*Wesley Brooks, Rebecca Carvin, Steve Corsi, Mike Fienen*

## 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 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  $\delta$  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 exceedances 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 concentration is below the regulatory standard and one when the concentration exceeds the standard. All of the binary modeling techniques herein use logistic regression (Hosmer Jr and Lemeshow 2004). Binary regression methods are indicated with a (b).

**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  $\delta$ . Any technique that was implemented with this weighting scheme was separately implemented without any weighting of the observations. The methods that weight the observations are indicated with a (w).

**Selection-only methods** The contest investigated whether certain modeling methods should be used only to select covariates. Once the predictor variables were selected, the regression model using those predictors were estimated using ordinary least squares for the continuous methods, or ordinary logistic regression for the binary methods. Selection-only methods are indicated by an (s).

## GBM

A gradient boosting machine (GBM) model is a so-called random forest model - a collection of many regression trees (Friedman 2001). Prediction is done by averaging the outputs of the trees. Two GBM-based techniques are explored - we refer to them as GBM-OOB and GBM-CV. The difference is in how the optimal number of trees is determined - GBM-CV selects the number of trees in a model using leave-one-out CV, while GBM-OOB uses the so-called out-of-bag 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 GBM-CV should more accurately estimate the prediction error.

All the GBM-OOB and GBM-CV models were estimated using the `gbm` package in R, using 10000 trees, shrinkage parameter 0.0005, minimum 5 observations per node, tree depth 5, and bagging fraction 0.5 [ @. ]

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 (AL) is a penalized regression method that simultaneously selects relevant predictors and estimates their coefficients (Zou 2006). The penalty part of an AL model is proportional to the sum of absolute values of the coefficients - covariates are dropped from the model when they incur a penalty larger than the amount by which they improve the model fit. For continuous response, AL estimates  $\hat{\beta}$  to minimize the penalized sum of squared residuals:

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

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

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

where, as for continuous response,  $\tilde{\beta}$  is a consistent estimate of the regression coefficients. For the contest,  $\gamma = 1$ , the  $\tilde{\beta}$  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 AL tuning parameter  $\lambda$  is selected to minimize the AICc (Hurvich, Simonoff, and Tsai 1998).

## Genetic algorithm

The genetic algorithm is a variable-selection method that works by analogy to natural selection, where so-called chromosomes represent regression models (Fogel 1998). 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.

## 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.

## 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 contains 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 predicted concentration by LOO CV and by LOYO CV. From the table, we can calculate the predictive error sum of squares (PRESS) and the area under the receiver operating characteristic (ROC) curve (AUROC), which are the statistics we use to summarize performance of the modeling methods.

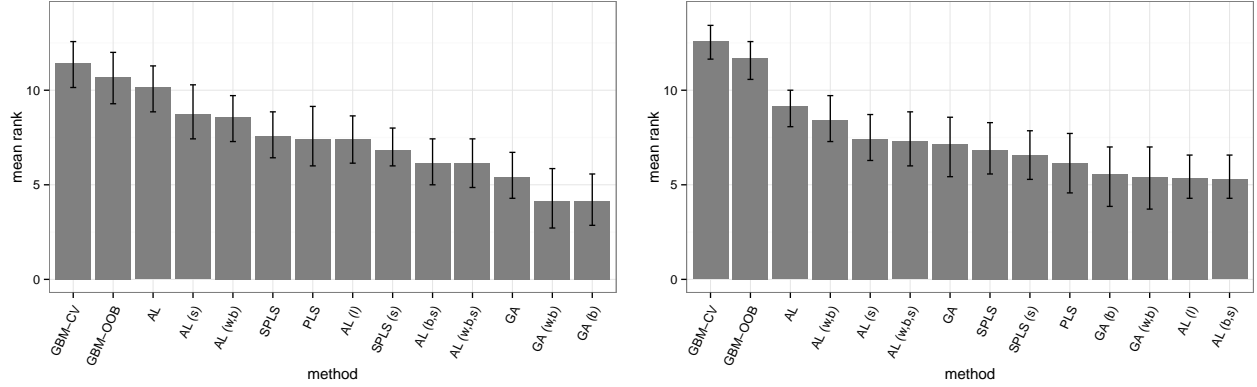
Row	Actual	PLS (LOO)	PLS (LOYO)	...	SPLS (LOO)	SPLS (LOYO)
1	2.54	2.35	2.22	...	2.29	2.55
2	2.59	1.87	1.79	...	1.91	1.23
⋮	⋮	⋮	⋮	...	⋮	⋮
166	1.57	1.93	2.06	...	1.83	2.07
167	3.38	1.84	2.01	...	1.80	1.71

In order to identify which modeling methods have the best performance across all sites, the summary statistics at each site were converted to ranks. We report the mean rank of each method across the sites. Uncertainty in the rankings is quantified by the bootstrap. Since PRESS and AUROC are functions of the results tables, the bootstrap procedure is carried out by resampling the rows of each results table and recalculating the ranks for each bootstrap sample. We used 1001 bootstrap samples of each results table in the analysis that follows.

## Results

### AUROC

The ROC curve is an assessment of how well predictions are sorted. The AUROC averages 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.



% latex table generated in R 3.1.1 by xtable 1.7-3 package % Thu Jul 31 17:37:16 2014

	GBM-OOB	AL	AL (s)	AL (w,b)	SPLS	PLS	AL (l)	SPLS (s)	AL (b,s)	AL (w,b,s)	GA
GBM-CV	0.86	0.84	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GBM-OOB		0.68	0.91	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00
AL			0.97	0.88	0.99	0.98	0.99	1.00	1.00	1.00	1.00

Table 2: Under leave-one-year-out cross validation, how often the mean AUROC rank of GBM-OOB, GBM-CV, or AL (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 % Thu Jul 31 17:37:16 2014

	GBM-OOB	AL	AL (w,b)	AL (s)	AL (w,b,s)	GA	SPLS	SPLS (s)	PLS	GA (b)	AL (b,s)
GBM-CV	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GBM-OOB		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AL			0.73	1.00	0.91	0.96	0.98	1.00	0.99	1.00	1.00

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

The mean LOO and LOYO ranks for all the methods are plotted in Figure [fig:auroc-boxplot]. The three top-ranked methods were GBM-CV, GBM-OOB, 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-OOB, GBM-CV, or the adaptive lasso exceeded each of the other modeling methods.

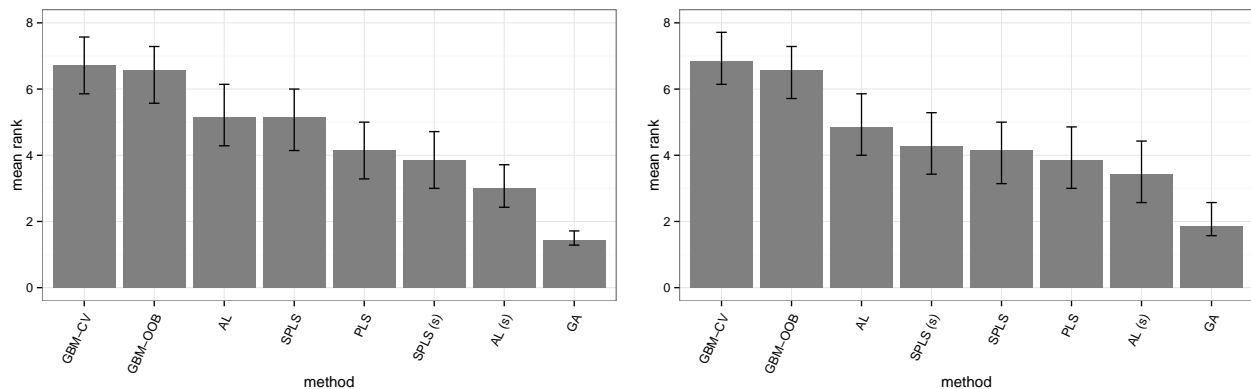
## PRESS

While AUROC quantifies how well a model sorts exceedances and non-exceedances, PRESS measures how accurately 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  $i = 1, \dots, n$  where  $n$  is the total number of predictions. Then PRESS computed as follows:

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

The PRESS statistic is of interest because a good model should accurately predict the bacterial concentration, but 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-CV, GBM-OOB, and the adaptive lasso. The pairwise comparison of modeling methods by PRESS are in Tables [table:auroc.pairs.annual] (for the leave-one-year-out analysis) and [table:auroc.pairs] (for the leave-one-out analysis).



% latex table generated in R 3.1.1 by xtable 1.7-3 package % Thu Jul 31 17:37:17 2014

	GBM-OOB	AL	SPLS	PLS	SPLS (s)	AL (s)	GA
GBM-CV	0.58	0.96	0.97	1.00	1.00	1.00	1.00
GBM-OOB		0.94	0.95	1.00	1.00	1.00	1.00
AL			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-OOB, GBM-CV, or AL (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 % Thu Jul 31 17:37:17 2014

## Narrowing the focus

By both AUROC and PRESS, and for both LOO and LOYO analyses, the three highest-ranked modeling methods were GBM-CV, GBM-OOB, and the adaptive lasso. The fourth-ranked method was not consistent

	GBM- OOB	AL	SPLS (s)	SPLS	PLS	AL (s)	GA
GBM-CV	0.65	1.00	1.00	1.00	1.00	1.00	1.00
GBM-OOB		0.97	0.99	1.00	1.00	1.00	1.00
AL			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-OOB, GBM-CV, or the adaptive lasso (in the rows) exceeded that of the other methods (in the columns).

across the different analyses. By the LOO CV analysis, the adaptive lasso was ranked better than the fourth-ranked method by AUROC, adalasso.weighted, on 73.027% of bootstrap samples and better than the fourth-ranked method by PRESS, spls.select, on 71.2288% of bootstrap samples. And by the LOYO CV analysis, the adaptive lasso was ranked better than the fourth-ranked method by AUROC, adapt.select, on 96.7033% of bootstrap samples and better than the fourth-ranked method by PRESS, spls, on 50.8492% of bootstrap samples.

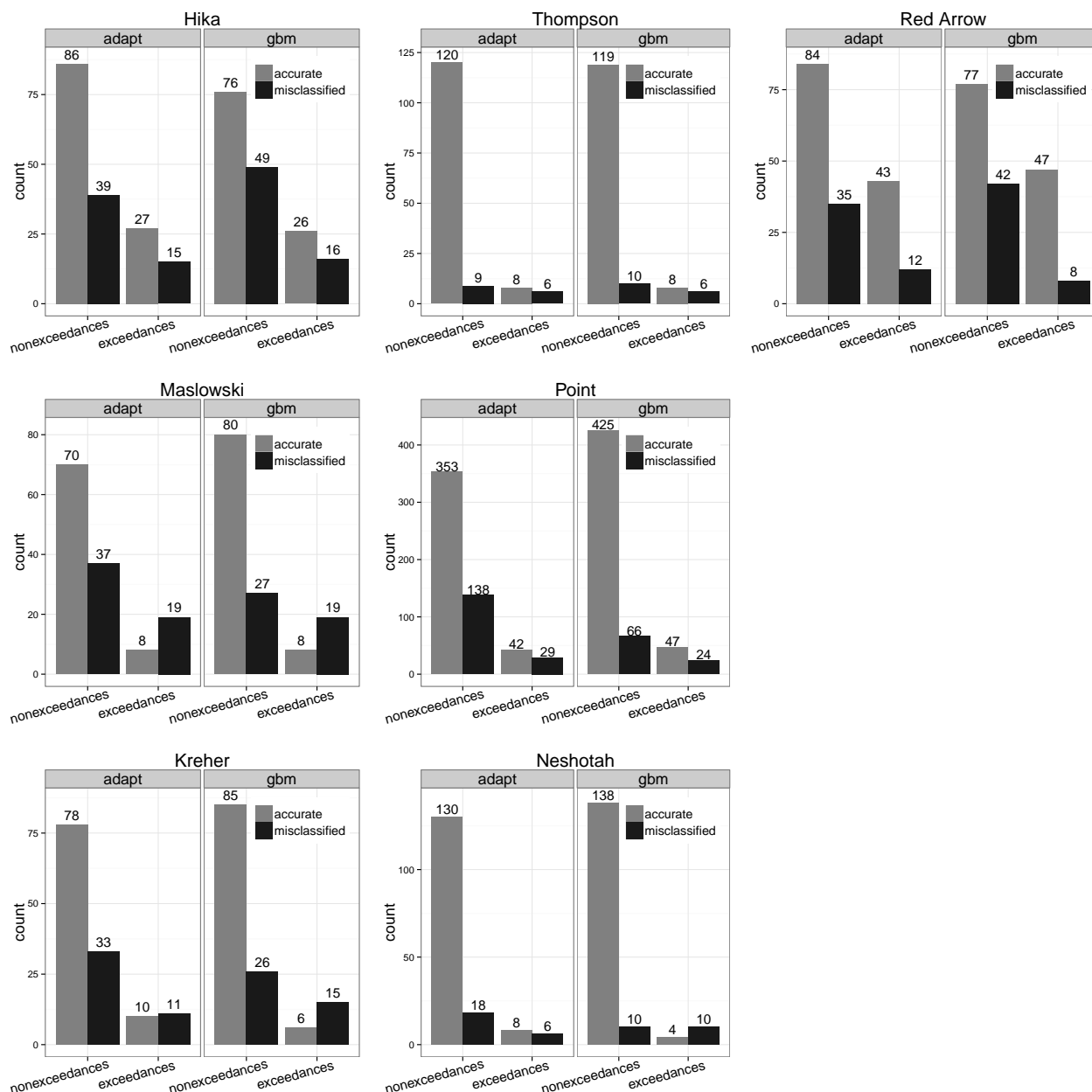
Therefore, we conclude that only the GBM methods and the adaptive lasso are worth considering further because they consistently outperform the other methods. We further narrow our study to GBM-OOB and the adaptive lasso because the GBM-OOB and GBM-CV methods are very similar and fitting a GBM-CV takes many times longer than a GBM-OOB model.

## Classification of responses

The real-world performance of any model for predicting exceedances will be measured by how well it distinguishes between exceedances and nonexceedances. This is quantified similar to AUROC except that AUROC is an average over all possible thresholds instead of an assessment of performance for a specific choice of threshold. Combined with a realistic carefully selected decision threshold, the LOYO CV results were used to simulate real-world use of the models.

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 exceedances in the training set.

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). In most cases, the counts are similar between the two techniques, with GBM-OOB tending to make a few more correct decisions. There are exceptions where adaptive lasso makes more correct decisions (e.g., Hika and Red Arrow).

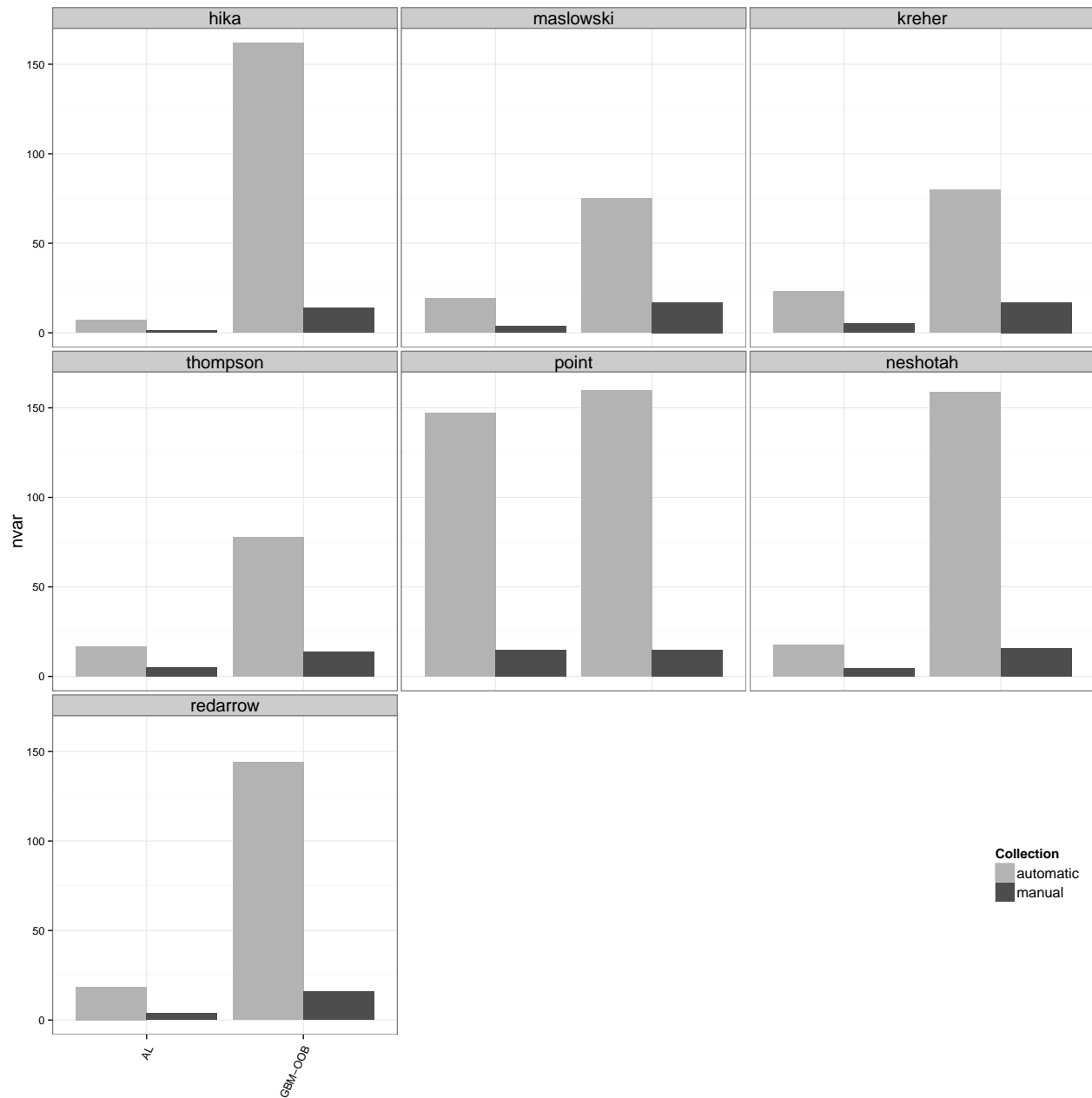


## Variable selection

It was noted in Section [Narrowing the focus](#) that GBM-OOB and the adaptive lasso are the two of the three best-ranked methods. One difference between the two is that the adaptive lasso does variable selection while GBM-OOB uses all of the available covariates. We look here at how many variables were used in the adaptive lasso models compared to the GBM-OOB models.

The variable counts are displayed in Figure [fig:vareselect-barchart]. At most of the sites, the adaptive lasso uses only a small fraction of the available covariates, but at Point the adaptive lasso uses almost all of the available covariates. That's because the variable selection criterion we used (AICc) is intended to minimize prediction error. As the amount of data increases, 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 an adaptive lasso model, and Point has far more observations than the other sites.





## Discussion

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-OOB versus GBM-CV → notes on how GBM-OOB 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*, 267–281. 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: 470–475.
- 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.
- Hosmer Jr, David W, and Stanley Lemeshow. 2004. *Applied Logistic Regression*. John Wiley & Sons.
- 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: 271–293.
- Wold, Svante, Michael Sjostrom, and Lennart Eriksson. 2001. “PLS-Regression: a Basic Tool of Chemometrics.” *Chemometrics and Intelligent Laboratory Systems* 58: 109–130.
- Zou, Hui. 2006. “The Adaptive Lasso and Its Oracle Properties.” *Journal of the American Statistical Association* 101: 1418–1429.