Beauty Contest

Wesley Brooks, Steve Corsi, Rebecca Carvin

July 25, 2014

Abstract

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

1 Introduction

2 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.

2.1 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

2.2 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).

2.3 Goals

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. The goal of modeling the bacterial concentration is to predict in advance when the concentration will exceed the limit. For the discussion to come, denote the regulatory standard by δ. Each model has two essential components: the mathematical model itself, $\mu(X, y)$, and a decision threshold, $\hat{\delta}$, that is used to interpret the model's predictions.

There is no reason that δ and $\hat{\delta}$ must be equal. Rather, each $\hat{\delta}$ should be chosen so that the number of false positives and false negatives are balanced to the satisfaction of the beach manager. Using cross validation allows us to set $\hat{\delta}$ to expect that the performance over future data will resemble what was observed over the testing data.

2.4 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 aplied to either continuous or binary regression and to either modeling, or variable selection only.

Continuous or 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:

$$z_i = I(y_i > \delta) \tag{2.1}$$

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.

$$g(p_i) = \log \frac{p_i}{1 - p_i} \tag{2.2}$$

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:

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

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.

2.4.1 GBM

GBM refers to the gradient boosting machine (GBM) of **?**. 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

2.4.2 Adaptive Lasso

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

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.

2.4.3 Genetic algorithm

The genetic algorithm [**?**] 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) **??**.

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 [**?**]. This study also investigates two genetic algorithm methods for logistic regression: one weighted (GA-logistic-weighted) and one unweighted (GA-logistic-unweighted).

2.4.4 PLS

Partial least squares (PLS) regression is a tool for building regression models with many covariates [**?**]. 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 **?**, using the PRESS statistic to select the number of components.

2.4.5 SPLS

Sparse PLS (SPLS) is introduced in **?**. The SPLS method combines the orthogonal decompositions of PLS with the sparsity of lasso-type variable selection. 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).

2.5 Implementation for beach regression

The response variable for our continuous regression models is the natural 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.Some pre-processing of the data was done. During pre-processing, some transformations were applied to the data. The beach water turbidity and the discharge of tributaries near each beach were log-transformed. Rainfall variables were all square-root transformed.

Include a table with pre/post processing discussion

2.6 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.

2.7 Performance Metrics

How did we evaluate the performance of each technique on all the different data sets

- AUROC: The area under the ROC (AUROC) curve assesses how accurately the predictions of the left-out observations are sorted.
- continuous variables using PRESS (skill –> Like Nash-Sutcliffe/R^2 over the fitted data)
- True/False Positives/Negatives (needs a threshold)
- Which variables are selected for models where variable reduction takes place
  - challenge regarding the fact that different variables are selected in each fold. Maybe use frequencies?
  - also the number of variables selected (metric of complexity)

Area under the ROC curve 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.

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:

$$\text{PRESS} = \sum_{i=1}^{n} \left(\tilde{y}_i - y_i \right)^2 . \tag{2.3}$$

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.

True positives, etc. While the AUROC is an important metric that should guide the choice of which modeling method to use, the real-world performance of any model for predicting exceedances will be measured not by its average performance over the possible range of thresholds, but by how well it distinguishes between exceedances and nonexceedances at the one specific threshold $\hat{\delta}$ that is used in the application. 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 $\hat{\delta}$.

Intuitively, $\hat{\delta}$ 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 $\hat{\delta}$ high to reflect this prior expectation. On the other hand, if the bacterial concentration often exceeds δ, the threshold $\hat{\delta}$ should be set lower in order to properly flag more of those exceedances. This intuition was encoded into how $\hat{\delta}$ was set for the LOYO models. Specifically, $\hat{\delta}$ was set to the qth quantile of the fitted values for non-exceedances in the training set, where $1 - q$ is the proportion of exceednaces in the training set. These LOYO predictions are summarized in the 2-by-2 tables of true positives, true negatives, false positives, and false negatives.

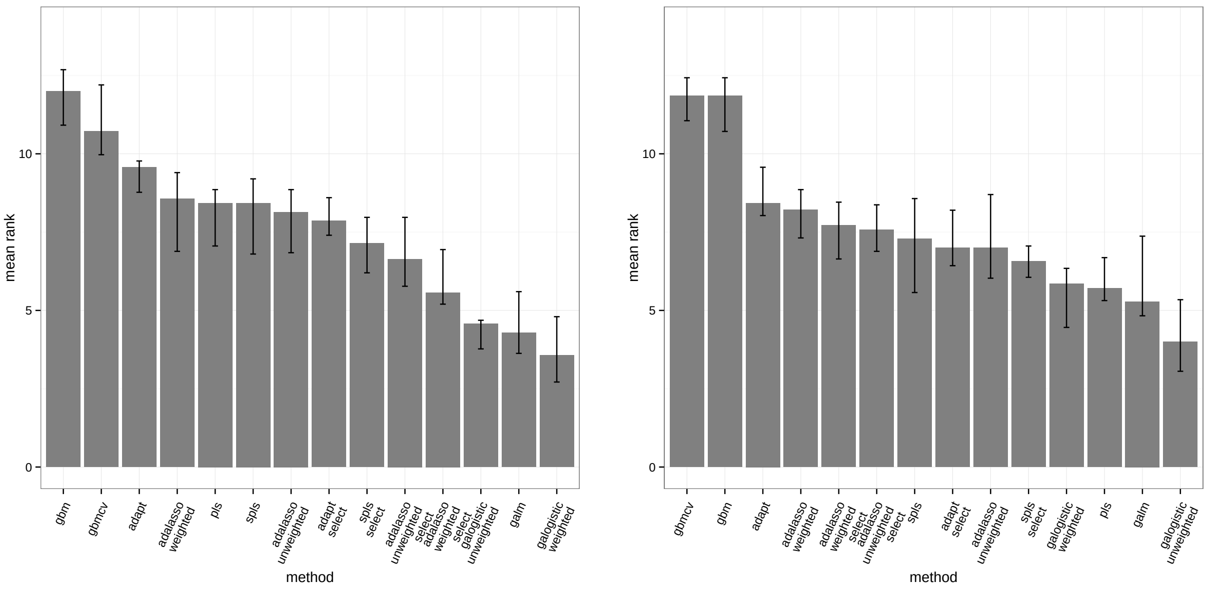
3 Results

3.1 Results tables

Results of the cross-validation procedures were compiled into tables

3.2 Comparing methods, and quantifying uncertainty in the ranks

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.



(a) LOO

(b) LOYO

Figure 3.1: Mean ranking of the methods by area under the ROC curce (AUROC) across all sites (higher is better). The error bars are 90% confidence intervals computed by the bootstrap. At left are the AUROC rankings from the leave-one-year-out cross validation, at right are the AUROC rankings from the leave-one-out cross validation.

In order to make a decision about which technique is best, 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.

3.3 AUROC

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 3.1. The top three techniques were GBM, GBMCV and adaptive lasso. The PLS and GALM techniques both appeared well down the list of the best techniques. In order to facilitate a pairwise comparison between modeling methods, Tables 1 (for the leave-one-year-out analysis) and 2 (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. For instance,

	gbmcv	adapt	adalasso weighted	pls	spls	adalasso unweighted	adapt select	spls select	adalasso unweighted select	adalasso weighted select	galm	galogistic unweighted	galogistic weighted
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	1.00
gbmcv		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			0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 1: 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).

	gbm	adapt	adalasso weighted	adalasso unweighted select	adalasso weighted select	adapt select	adalasso unweighted	spls	spls select	pls	galm	galogistic weighted	galogistic unweighted
gbmcv	0.60	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			0.60	0.80	0.80	1.00	0.80	0.80	1.00	1.00	1.00	1.00	1.00

Table 2: 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).

3.4 PRESS

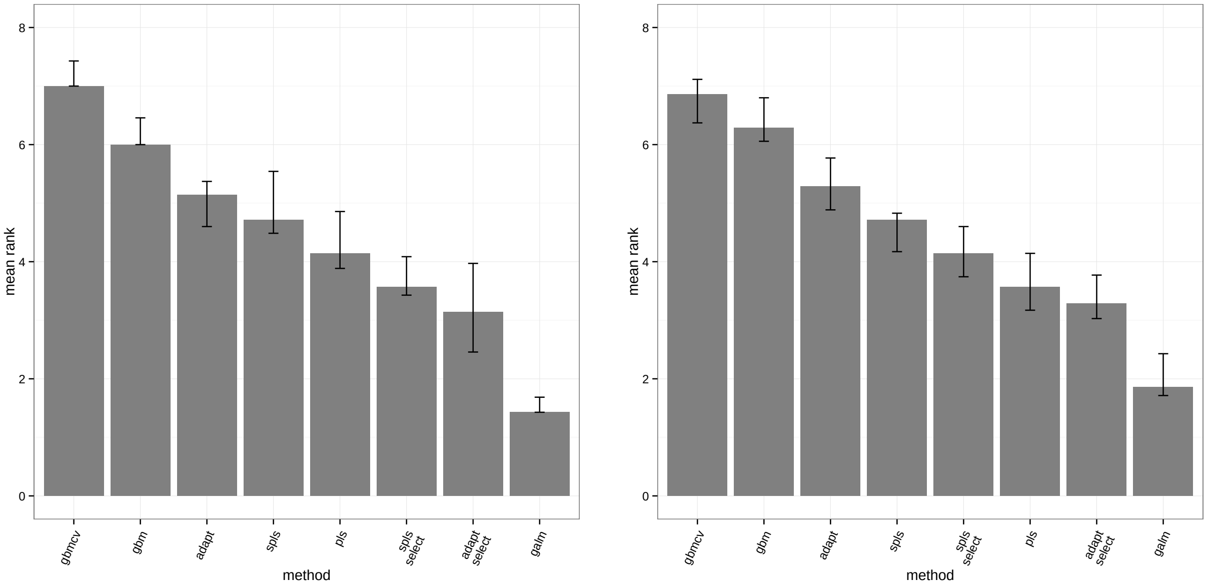
The methods are ranked at each site by PRESS and a mean rank (across sites) is computed for each method. The mean LOO and LOYO ranks are plotted in Figure 3.2. The top three techniques were GBM, GBMCV and adaptive lasso. The PLS and GALM techniques both appeared well down the list of the best techniques.

	gbm	adapt	spls	pls	spls select	adapt select	galm
gbmcv	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
adapt			0.60	0.80	1.00	1.00	1.00

Table 3: 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).

3.5 Classifying the models' decisions

Since GBM, GBMCV, 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, $\hat{\delta}$, 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



(a) LOO

(b) LOYO

Figure 3.2: Mean ranking of the methods by predictive error sum of squares (PRESS) across all sites (higher is better). The error bars are 90% confidence intervals computed by the bootstrap. At left are the PRESS rankings from the leave-one-year-out cross validation, at right are the PRESS rankings from the leave-one-out cross validation.

	gbm	adapt	spls	spls select	pls	adapt select	galm
gbmcv	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
adapt			1.00	1.00	1.00	1.00	1.00

Table 4: 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).

beaches. In Figure 3.3, 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 GBMCV methods are so similar except that GBMCV takes much longer to run. 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).

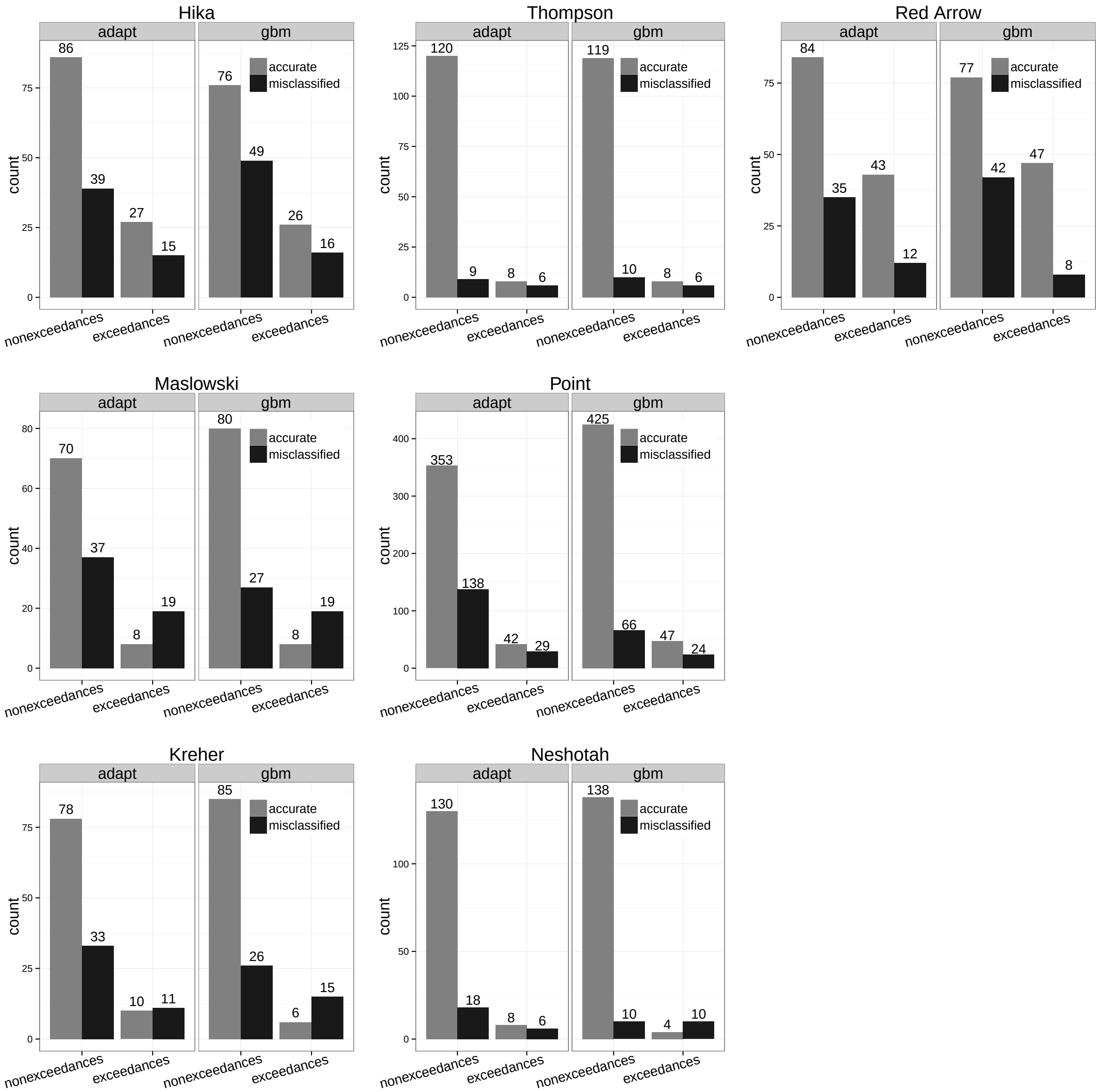


Figure 3.3: Mean ranks of the modeling techniques across the seven sites. At left are the mean ranks under leave-one-out cross validation, at the right are the mean ranks from leave-one-year-out cross validation.

4 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)

5 Acknowledgments

6 References