Comparing methods for predicting health advisories for beach water

Wesley Brooks, Rebecca Carvin, Steve Corsi, Mike Fienen

COMMENT: General comments: 1. Very nice work here. The writing is concise and clear. The organization is well done. Most comments are just minor issues or some things that might help with clarification. 2. Need to be a bit more consistent with acronyms. Once you define an acronym, use it throughout. There are cases where the full spelling and the acronym are mixed throughout the manuscript. 3. Some of the table and figure references are muddled up in the linking process. 4. After all is complete, I wondered if we should include reference to virtual beach and the methods that are included earlier in the manuscript. I am not convinced either way yet, but it would be worth a little discussion as to where it might be appropriate. Maybe a mention in the methods when they are being described? It might also be worth mentioning that VB only had OLS/GA options until recently. This would fit in the intro section where it is mentioned that OLS is the most common method, and would serve to strengthen that statement.

1 Abstract

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

2 Introduction

Fecal indicator bacteria (FIB) in beach water are often used to indicate contamination by harmful pathogens (Cabelli et al.; T. J. Wade et al.; Timothy J. Wade et al.; Fleisher et al.). The United States Environmental Protection Agency (USEPA) has established, through epidemiological studies, that FIB concentration is associated with human health outcomes (Cabelli; Dufour; USEPA 1986). Accordingly, the state of Wisconsin has established regulatory standards for beach water quality, stating that a beach should be posted with a swimmer's advisory when the concentration of the FIB *Escherichia coli* exceeds 235 colony forming units (CFU) / 100 mL (USEPA 2012; WDNR 2012). Traditional analysis methods for FIB concentration requires $18 - 24$ hours for culturing a sample, so the decision to post an advisory is often made based on the previous day's FIB concentration, which is the so-called "persistence model" for beach management (USEPA). Previous research has shown that the concentration of FIB in beach water can vary substantially during the $18 - 24$ h analysis period, with the result that the persistence model often provides incorrect information for posting warnings (Whitman et al.; R. L. Whitman and Nevers). Thus, at beaches managed using the persistence model, the public is sometimes exposed to health risks or unnecessarily deprived of recreation opportunities.

In order to have more immediate knowledge of the FIB concentration, it is now common to use regression models that "nowcast" the FIB concentration based on some easily observed surrogate covariates, e.g. turbidity and running 24 h rainfall total (Brandt et al.; Olyphant and Whitman). Numerous regression techniques have been used to generate nowcast models of FIB concentration. The techniques include ordinary least squares (OLS) (Nevers and Whitman; Francy and Darner), partial least squares (PLS) (Hou, Rabinovici, and Boehm; Brooks, Fienen, and Corsi 2013), logistic regression (Waschbusch et al.; Jin and Englande Jr.), decision trees (Stidson, Gray, and McPhail 2012), random forests (Parkhurst et al.; Jones, Liu, and Dorovitch 2012), and artificial neural networks (Kashefipour, Lin, and Falconer 2005; He and He). A thorough review of the regression techniques being used in nowcast models for FIB concentration is provided by Brauwere, Ouattara, and Servais (2014). An assessment of seven methods of regression for FIB concentration in beach water at Santa Monica Beach in California identified classification trees, artificial neural networks, and logistic regression as the three best methods (Thoe et al. 2014).

Ordinary least squares regression is the most commonly used regression technique in the nowcast models (Brauwere, Ouattara, and Servais 2014). However, OLS is well-known for drawbacks like overfitting, difficulty of variable selection, and the inflexibility of its linear modeling structure (Ge and Frick). The literature suggests that many regression techniques have been successfully used for nowcast modeling, but due to differences in such factors as local conditions, data handling, and performance validation, it is not possible to identify the best regression technique for nowcast modeling by comparing different models at different sites. In this study, fourteen regression techniques are evaluated in nowcast models at seven Wisconsin beaches with four years of data. The results are compared to identify the techniques that more accurately predict instances when a swimmer's advisory should be posted. This comparison is designed to provide insights that may be lost when comparing individual methods at single sites.

The remainder of the paper is organized as follows: in the next section we discuss data collection and handling, describe the regression techniques, and explain how the comparisons were made. Next, we present the results of comparing the methods by several metrics including: area under the ROC curve; predictive error sum of squares; and raw number of correct/incorrect predictions. Finally, we discuss what the comparison suggests about which are the best choices for a regression technique in a nowcast model.

3 Data

The seven beach sites analyzed in this study are located within two distinct regions of Wisconsin. Three of the sites are on Chequamegon Bay (part of Lake Superior) and the remaining five are in Manitowoc County on Lake Michigan. For each site in the study, the data used to estimate the predictive models for FIB concentration were measured by a combination of automatic sensing and manual sampling.

3.1 Site descriptions

3.1.1 Chequamegon Bay sites

Chequamegon Bay is approximately 12 miles long and ranges from two to six miles in width, with a maximum depth of 35 feet. The three Chequamegon Bay/Lake Superior beaches are influenced by nearby streams, as well as by urban runoff from Ashland and Washburn, Wisconsin. Thompson beach is within the small town of Washburn, on the north side of the bay. Next to the beach are a playground, RV campsites, piers and boat launch. There are two flowing (artesian) wells that drain to the beach. Thompson Creek, about 1000 ft from the beach, is the nearest stream. Maslowski beach is on the west side of Ashland, on the south side of the Chequamegon Bay. A playground and parking area are near the beach. Two flowing wells are near the swim area, and Fish Creek (one mile west of the beach) and Whittlesley Creek (two miles northwest) are the nearest streams. Kreher beach is in Ashland, 2.5 mi northeast of Maslowski beach. Kreher Park has an RV campground, playground and boat launch, and is nearest to Bay City Creek, which is 0.5 mi east of the beach. All of the listed streams are influenced by areas of agricultural and forested land use, with Bay City Creek also influenced by urban land use (Francy et al. 2013). Contributions from Fish Creek are dynamic due to a wetland at the creek's outlet that is influenced by the lake level. There were no real-time discharge measurements available for Fish Creek.

3.1.2 Manitowoc County sites

Red Arrow beach is within the city of Manitowoc. It has numerous potential influences on water quality, including the mouth of the Manitowoc River one mile north and urban runoff draining to the beach through storm sewer outlets. The Manitowoc River is dominated by agricultural land use, but there is some urban influence from the city of Manitowoc. The Manitowoc sewage treatment plant sits at the mouth of the Manitowoc River. Neshotah beach is in the small community of Two Rivers. Small storm sewers drain to the north and to the south directly adjacent to the beach boundaries, and the mouth of the Twin River is

0.5 mi south of the beach. The Twin River drains an agricultural watershed. Point Beach State Park is approximately 11 mi north of Manitowoc, about 2.5 mi north of the mouth of Molash Creek whose watershed encompasses a mix of agricultural land use and wetland area. The mouth of Twin River is 6.3 mi south and the mouth of the Kewaunee River is 16 mi north of Point Beach State Park. The Kewaunee River is also dominated by agricultural land use. Hika beach is south of the city of Manitowoc near the small community of Cleveland. Large floating mats of *Cladophora* algae are common. Centerville Creek, a small stream dominated by agricultural land use, drains into the lake adjacent to the beach.

3.2 Data sources

Data collection and sample analysis followed methods described in Francy et al. (2013). Concentration of *E. coli* was measured at each beach $2-4$ times per week for $12-14$ weeks between Memorial Day and Labor Day from 2010 through 2013. Samples were collected from the center of the length of the beach, 12 inches below the water surface where total water depth was 24 inches. All samples were quantified by use of the Colilert® QuantiTray/2000 method, which is reported as the most probable number (MPN) of *E. coli* colony forming units (CFU) and is read after 24 hours of incubation ("Colilert Test Kit Procedure").

Covariates were compiled from a variety of sources including online data and manual measurements. Online data were accessed using Environmental Data Discovery and Transformation (EnDDaT), a web service that accesses data from a variety of sources, compiles and processes the data, and performs common transformations (U.S. Geological Survey 2014). Three sources of data were accessed: National Water Information System (NWIS), North Central River Forecasting Center (NCRFS), and Great Lakes Costal Forecasting System (GLCFS). Variables acquired through these sources included: river discharge, precipitation, lake current vectors, wave height, wave direction, lake level, water temperature, air temperature, wind vector, and percent cloud cover.

Most covariates from online sources were available in hourly increments with the exception of NWIS data which were available in 15 minute increments. In order to make best use of this high-frequency data for daily predictions, several summary statistics were calculated over several time windows for use as potential covariates. The use of 1, 2, 6, 12, 24, 48, 72, and 120 hour time windows for calculating the summary statistics follows recent research showing that selecting from windowed and lagged versions of raw high-frequency covariates can improve the predictive accuracy of regression models (Richard Zepp 2012). The choice of summary statistics to include as potential covariates was guided by scientific judgement regarding phenomena that could affect the FIB concentration. For example, standard deviation of water temperature measurements over the window period reflected the variability in water temperature, which may affect the survival and growth of FIB; the sum of rainfall measurements over the window period indicated the magnitude of recent rain events, which may be associated with FIB washed into the lake from sources on land; and the mean of cloud cover measurements over the window period may measure the degree to which UV light was inhibited from breaking down FIB colonies in the water. The available summary statistics from EnDDaT were the mean, minimum, maximum, difference, sum, and standard deviation.

Manually observed data were instantaneous observations that had the benefit of being measured when and where the FIB samples were collected. However, these covariates were measured only once per day and at greater expense than the online data because the data had to be collected by field personnel. Manual data collection was guided by the USEPA's Great Lakes Beach Sanitary Survey (USEPA 2008). Among the manually measured data were turbidity, wave height, number of birds present, number of people present, amount of algae floating in the swim area and on the beach, specific conductance, water and air temperature, wind direction, and wind speed. Every beach dataset included turbidity, but other field variables occasionally had to be dropped from some of the datasets because of missing values or questionable reliability.

3.3 Data transformations

The response for our continuous regression models is the base-10 logarithm of the FIB concentration. For the binary regression models, the response is an indicator of whether the concentration exceeds the BAV.

Transformations were applied to some of the covariates 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. These transformations were based on the performance of previous studies and were applied to all datasets (Ge and Frick; Frick, Ge, and Zepp).

4 Methods

4.1 Definitions

Fo each site, let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the vector of $\log_1 0$ FIB concentration measurements, let n be the number of observations, and let p be the number of explanatory covariates. The beach action value (BAV) of 235 CFU / 100 mL was recommended by the USEPA as the “do not exceed” threshold in order to limit gastrointestinal illnesses among those coming into contact with beach water to 36 cases per 1000 people (USEPA 2012). The BAV is represented symbolically in equations by δ. Define an exceedance as a meansured FIB concentration that exceeds the BAV. Conversely, a nonexceedance is a measured FIB concentration that does not exceed the BAV.

Predictions are the result of applying a model to data that was not used to estimate the model. The predicted $\log_1 0$ FIB concentration is denoted by a tilde (e.g., $\tilde{y}_i$). On the other hand, applying the model to the same data as was used to estimate the model produces fitted values, which are denoted by a hat (e.g., $\hat{y}_j$). Define a predicted exceedance as when a model predicts that the FIB concentration exceeds the BAV. This is not the same as $\tilde{y}_i > \delta$ because predictions are compared to a decision threshold $\hat{\delta}$ rather than to the BAV δ. The decision threshold $\hat{\delta}$ is a parameter that can be adjusted to tune the predictive performance. For instance, increasing the decision threshold reduces the number of false positives but increases the number of false negatives. Setting the decision threshold is an important detail that is discussed in Section 5.4.

4.2 Listing of statistical techniques

Fourteen different regression modeling techniques were considered (Table 1). Each technique uses one of five modeling algorithms: the gradient boosting machine (GBM), the adaptive Lasso (AL), the genetic algorithm (GA), partial least squares (PLS), or sparse PLS (SPLS). Each technique is applied to either continuous or binary regression and to either variable selection and model estimation, or variable selection only.

4.2.1 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:

$$z_i = \begin{cases} I\left(\tilde{y}_i < \delta\right) = 0 \\ I\left(\tilde{y}_i \geq \delta\right) = 1 \end{cases}$$

where $\tilde{y}_i$ is the predicted concentration. 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 (Hosmer Jr and Lemeshow 2004). Binary regression methods are indicated with a (b).

4.2.2 Weighting of observations in binary regression

The concentration of FIB in the water at a single beach on a single day can be subject to a large degree of spatiotemporal heterogeneity (R. L. Whitman and Nevers). Thus, when the concentration in a sample is observed to fall near the BAV, there is considerable uncertainty as to whether an independent sample from the same date and location would or would not exceed the BAV. A weighting scheme for the binary regression techniques was designed to reflect this ambiguity by giving more weight to observations far from the BAV. In the weighting scheme, observations were given weights w_i for $i = 1, \ldots, n$, where

$$w_i = (y_i - \delta)/\hat{\mathrm{sd}}(y)$$

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

$$\bar{y} = \sum_{i=1}^{n} y_i/n.$$

That is, the weights are equal to the number of standard deviations that the observed concentration lies from the regulatory threshold. Any technique that was implemented with this weighting scheme was separately implemented without any weighting of the observations. The methods using the weighting scheme are indicated by (w).

4.2.3 Selection-only methods

The contest investigated whether certain modeling methods should be used only to select covariates. Once the covariates were selected, the regression model using those covariates was 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).

4.2.4 Listing of modeling algorithms

4.2.4.1 GBM A GBM model is a so-called random forest model - a collection of many regression trees, each fitted to a randomly drawn subsample of the training data (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 cross validation (CV), while GBM-OOB uses the so-called out-of-bag error estimate, where the predictive error of each tree is estimated by its predictive error over the observations that were left out when fitting the tree. In contrast, the predictive error of CV is estimated from observations that are left out from the training data altogether, and are therefore not used in the fitting of any trees. 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). However, GBM-CV should more accurately estimate the prediction error.

4.2.4.2 Adaptive Lasso The least absolute shrinkage and selection operator (Lasso) is a penalized regression method that simultaneously selects relevant covariates and estimates their coefficients (Tibshirani 1996). The AL is a refinement of the Lasso that possesses the so-called “oracle” properties of asymptotically selecting exactly the correct covariates and estimating them as accurately as would be possible if their identities were known in advance (Zou 2006). To use the AL for prediction requires selecting a tuning parameter. For the contest, the AL tuning parameter λ is selected to minimize the corrected Akaike Information Criterion (AICc) (Akaike 1973; Hurvich, Simonoff, and Tsai 1998).

4.2.4.3 Genetic algorithm

Here, the GA is used to select variables for either an OLS or a logistic regression model. By analogy to natural selection, so-called chromosomes in the GA represent regression models (Fogel 1998). A covariate 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 AICc.

4.2.4.4 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 covariates 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, whereas 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 (Brooks, Fienen, and Corsi 2013), using the PRESS statistic to select the number of components.

4.2.4.5 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.

Name	Algorithm	Binary	Weighted	Selection Only
GBM-OOB	Gradient boosting			
GBM-CV	Gradient boosting			
AL	Adaptive Lasso			
AL (s)	Adaptive Lasso			X
AL (b)	Adaptive Lasso	X		
AL (b,w)	Adaptive Lasso	X	X	
AL (s,b)	Adaptive Lasso	X		X
AL (s,b,w)	Adaptive Lasso	X	X	X
GA	Genetic algorithm			
GA (b)	Genetic algorithm	X		
GA (b,w)	Genetic algorithm	X	X	
PLS	Patrial least squares			
SPLS	Sparse partial least squares			
SPLS (s)	Sparse partial least squares			X

Table 1: Comprehensive list of the modeling methods analyzed in this study. Listed for each method are the method's abbreviation, the algorithm used by the method, and indicators of whether the method

4.3 Cross Validation

Our assessment of the modeling techniques is based on their performance in predicting exceedances of the BAV. Two types of cross validation were 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 result of that held out observation, and the process - including estimating a new predictive model - is repeated for each observation. On the other hand, 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 CV internally to select tuning parameters. In those cases the internal CV was conducted by subdividing the model data, and never looking at the held-out observation(s). This process was separate from - and did not affect - the CV to assess predictive performance.

4.4 Comparing methods, and quantifying uncertainty in the ranks

Results were compiled into one table for each site where each observation corresponds to a row in the table. For example, a few rows from the results table at Hika are presented in Figure 1. The results table has a column for the observed $\log_{10}$ FIB concentration and, for each method, columns for the predicted concentration by LOO CV and by LOYO CV. From the table, performance of the modeling methods was summarized by calculating the predictive error sum of squares (PRESS) and the area under the receiver operating characteristic (ROC) curve (AUROC).

Row	$\log_{10}$ FIB	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

Table 2: An example of how the results for a site (Hika here) were compiled into a results table. The summary statistics used to compare predictive performance (area under the ROC curve and predictive error sum of squares) were calculated from the table. Confidence intervals for the summary statistics were computed via the bootstrap by resampling (with replacement) the rows of the results table.

To identify which modeling methods had the best performance across all sites, the methods at each site were ranked from worst to best according to the performance summary statistics (the ranks were taken worst to best so that larger numbers represent better performance). The mean rank of each method was then taken across the sites as an measurement of how each of our modeling methods performed relative to the others. 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 $1,000$ bootstrap samples of each results table in the analysis that follows.

5 Results

5.1 AUROC

The ROC curve is an assessment of how well predictions were separated into exceedances and nonexceedances (Hanley and McNeil 1982). Every possible value of the decision threshold $\hat{\delta}$ corresponds to the point on the ROC curve with coordinates $(1 - \text{specificity}, \text{sensitivity})$, where the specificity and sensitivity are defined as

$$\text{specificity}(\hat{\delta}) = \sum_{i=1}^{n} I(\tilde{y}_i \leq \hat{\delta}) I(y_i \leq \delta) / \sum_{j=1}^{n} I(y_j \leq \delta)$$
$$\text{sensitivity}(\hat{\delta}) = \sum_{i=1}^{n} I(\tilde{y}_i > \hat{\delta}) I(\tilde{y}_i > \delta) / \sum_{j=1}^{n} I(y_j > \delta).$$

The AUROC averages the model's performance over the range of possible thresholds. A model which perfectly separates exceedances from non-exceedances in prediction would have an AUROC of one, while a model that predicts exceedances no better than a coin flip would have an expected an AUROC of 0.5.

COMMENT: The ranks are opposite of what I typically consider as ranks–usually I would consider 1 to be the best ranking. It is clear enough in the title that higher is better and certainly makes more sense visually to see the best method with the largest bar. Interesting that the lowest ranks are between 4 and 5, indicating that the lower ~2/3 ranked methods change places a fair bit

COMMENT: These two tables could be made into one two-section table to save a bit of space. They do logically go together well.

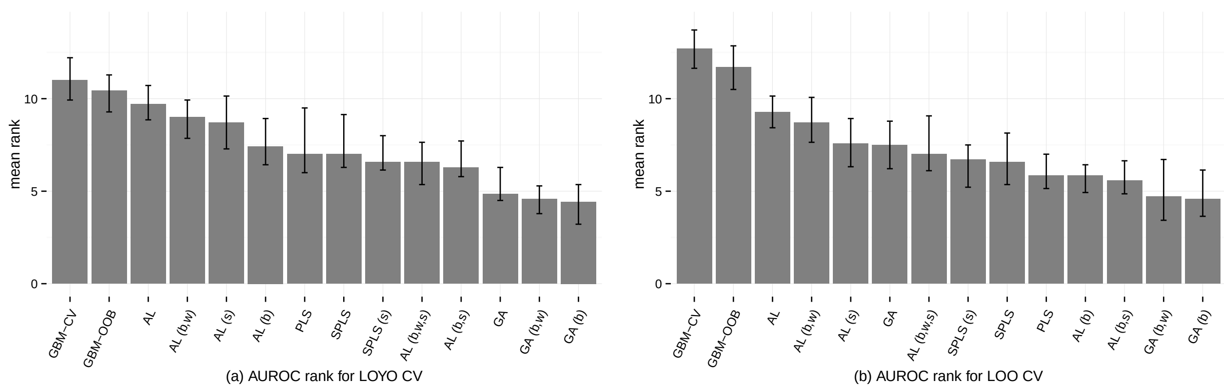

Figure 1: Mean ranking of the methods across all seven sites by area under the receiver operating characteristic curve (AUROC) and predictive residual sum of squares (PRESS) (higher is better). The error bars are 90% confidence intervals computed by the bootstrap. At left are the AUROC rankings from the leave-one-year-out cross validation (a), at right are the AUROC rankings from the leave-one-out cross validation (b)

The mean LOO and LOYO ranks for all the methods are plotted in Figure 1. The three top-ranked methods were GBM-CV, GBM-OOB, and AL. In order to facilitate a pairwise comparison between modeling methods, Tables 2 (for the leave-one-year-out analysis) and 3 (for the leave-one-out analysis) show the frequency that the mean AUROC rank of GBM-OOB, GBM-CV, or AL exceeded each of the other modeling methods.

5.2 PRESS

While AUROC quantifies how well a model sorts exceedances and non-exceedances, PRESS measures how accurately a model's predictions match the observed FIB concentration. The PRESS can only be computed

Leave-one-year-out cross-validation:

	GBM-OOB	AL	AL (b,w)	AL (s)	AL (b)	PLS	SPLS	SPLS (s)	AL (b,s)	AL (b,w,s)	GA	GA (b,w)	GA (b)
GBM-CV	0.82	0.82	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GBM-OOB		0.73	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AL			0.73	0.91	1.00	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Leave-one-out cross-validation:

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

Table 3: Under leave-one-year-out (top) and leave-one-out (bottom) cross validation, frequency of the mean AUROC rank of GBM-OOB, GBM-CV, or AL (in the rows) exceeding that of the other methods (in the columns).

for continuous regression methods. Recalling that the ith observed $\log_1 0$ FIB concentration is denoted y_i and that the corresponding prediction is denoted $\tilde{y}_i$ for $i = 1, \dots, n$ where n is the total number of predictions, the PRESS is given by

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

The PRESS statistic is of interest because a good model should accurately predict the bacterial concentration, but for assessing regression models for FIB concentration, AUROC is more important than PRESS because it directly measures the models' abilities to distinguish exceedances from nonexceedances. That said, we expect the two statistics to usually agree on which modeling methods perform best.

The rankings of the methods by PRESS are plotted in Figure 2. The top three techniques under both LOO and LOYO analysis were GBM-CV, GBM-OOB, and AL. The pairwise comparison of modeling methods by PRESS are in Tables 4 (for the leave-one-year-out analysis) and 5 (for the leave-one-out analysis).

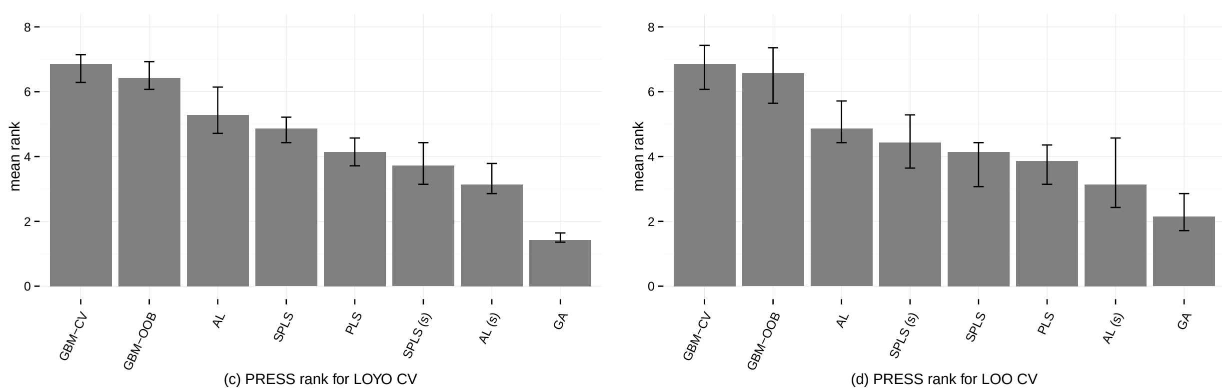

Figure 2: Mean ranking of the methods by predictive error sum of squares (PRESS) across all sites (higher is better). The error bars are 90% confidence intervals computed by the bootstrap. At left are the PRESS rankings from the leave-one-year-out cross validation (a), at right are the PRESS rankings from the leave-one-out cross validation (b).

5.3 Narrowing the focus

*COMMENT:*Probably just include the abbreviations for the models that are referenced beyond the 1-3 ranked methods. The abbrevs are used in tables and figs and that is what the reader is used to by this point in the

Leave-one-year-out cross-validation:

	GBM-OOB	AL	SPLS	PLS	SPLS (s)	AL (s)	GA
GBM-CV	0.73	0.91	1.00	1.00	1.00	1.00	1.00
GBM-OOB		0.91	1.00	1.00	1.00	1.00	1.00
AL			0.73	0.91	1.00	1.00	1.00

Leave-one-out cross-validation:

	GBM-OOB	AL	SPLS (s)	SPLS	PLS	AL (s)	GA
GBM-CV	0.55	1.00	1.00	1.00	1.00	1.00	1.00
GBM-OOB		1.00	0.91	1.00	1.00	1.00	1.00
AL			0.82	0.91	1.00	1.00	1.00

Table 4: Under leave-one-year-out (top) or leave-one-out (bottom) cross validation, frequency of the mean PRESS rank of GBM-OOB, GBM-CV, or AL (in the rows) exceeding that of the other methods (in the columns).

manuscript. Also, the acronyms should be upper case (SPLS was in lower case).

By both AUROC and PRESS, and for both LOO and LOYO analyses, the three highest-ranked modeling methods were GBM-CV, GBM-OOB, and AL. The fourth-ranked method was not consistent across the different analyses. By the LOO CV analysis, AL was ranked better than the fourth-ranked method by AUROC, AL (b,w), on 72.7% of bootstrap samples and better than the fourth-ranked method by PRESS, SPLS (s), on 81.8% of bootstrap samples. And by the LOYO CV analysis, AL was ranked better than the fourth-ranked method by AUROC, AL (b,w), on 72.7% of bootstrap samples and better than the fourth-ranked method by PRESS, SPLS, on 72.7% of bootstrap samples.

Therefore, we consider only the GBM methods and AL for the following analyses because they consistently outperform the other methods. We further narrow our study to GBM-OOB and AL because the GBM-OOB and GBM-CV methods showed similar performance but fitting a GBM-CV takes many times longer than a GBM-OOB model. While we focus on the AL and GBM-OOB . . . SOMETHING HERE ABOUT STILL LOOKING AT OTHERS???

5.4 Classification of responses, and the decision threshold

COMMENT: I like this figure–nicely done. This fig should be placed in the results section rather than the discussion section. Need to work out the figure numbers in the text. Something is going wrong with the links.

In real-world use, a model's performance will be judged by how well it distinguishes between exceedances and nonexceedances. While AUROC measures how well exceedances and nonexcedances are sorted among the predictions, AUROC is an average accuracy over all possible thresholds. In order to provide the assessment most relevant for real-world use, the LOYO CV results were used to simulate how many correct and incorrect predictions would be seen from an AL, GBM-OOB, or GBM-CV model with a specific choice of decision threshold. Using the LOYO CV results simulates the common scenario that a model is estimated at the beginning of each beach season and used to make predictions during that season, with a new model incorporating the new season of data estimated the following year into the new model's training data.

Intuitively, the decision threshold should adapt to the conditions that are observed in the beach's training data. If, for instance, exceedances were rare in the training data, then we expect few exceedances in the future, and should set the decision threshold high to reflect this 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 $\hat{\delta}$ was set to the q^{th} quantile of the fitted values of non-exceedances in the training set, where $1 - q$ was the proportion of training set observations that were exceedances.

In Figure [fig:counts-barcharts], we look at the counts on a per-beach basis of four categories of decisions: true negatives (correct predictions of nonexceedances), false positives (incorrect predictions of exceedances) true positives (correct predictions of exceedances), and false negatives (incorrect predictions of nonexceedances). In most cases, the counts were similar between the three techniques, with GBM-OOB and GBM-CV both tending to make a few more correct decisions than AL. There are, however, exceptions where AL made more correct decisions (e.g., Hika and Red Arrow).

5.5 Variable selection

COMMENT: Fig needs site labels. It would also be reasonable to use the same colors (greyscale) as you do in the previous barchart for consistency. This fig should be placed in the results section rather than the discussion section.

It was noted in Section Narrowing the focus that GBM-OOB and AL are two of the three best-ranked methods. One difference between the two is that AL does variable selection while GBM-OOB uses all of the available covariates. We explore here how many covariates were used in AL models compared to the GBM-OOB models.

The covariate counts are displayed in Figure [fig:varselect-barchart]. At most of the sites, AL uses only a small fraction of the available covariates, but at Point, AL uses almost all of the available covariates. This is due to the variable selection criterion we used (AICc) which is intended to minimize prediction error. As the amount of data increases, we accumulate enough information 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 AL model, and Point has far more observations than the other sites.

6 Discussion

The GBM-CV, GBM-OOB, and AL methods showed the best results by both PRESS and AUROC, under LOO and LOYO cross validation. Though GBM-CV was a bit more accurate than GBM-OOB in all the settings, the small improvement in accuracy may not outweight the large additional cost in time to fit the model. However, the additional computational cost is incurred only once when the model is estimated - given a new observation of beach data, both the GBM-CV and GBM-OOB models produce predictions nigh-instantaneously. Where predictive accuracy is the most important consideration and no difficulty is anticipated in acquiring the data, it is hard to argue against using a GBM-type model.

The predictive performance of the AL models was somewhat worse than that of the GBM models, but by including a variable selection step, the AL models reduce the number of covariates that must be measured in order to make daily predictions. A model that requires fewer covariates is less expensive and more robust (as the probability of encountering some missing data increases with the number of required covariates). This is particularly important for manually-collected covariates because collecting data by hand takes more time and costs more money than accessing publically available data from a web service. Across all the sites, the ratio of manually-collected to automatically collected covariates in the AL models seems to mirror the ratio among all available covariates, indicating that neither the manually- nor automatically-collected covariates are systematically more important to predicting the bacterial concentration. Some covariates tended to appear at every site in the AL models (and other models that include a variable selection step). The manually-collected covariates that were consistently selected for the models were the (log) turbidity in the beach water, and wave height at the beach.

Where minimizing the number of covariates is important, the selection criterion used here (corrected AIC) may not be appropriate. In that case, the Bayesian information criterion (BIC) is more parsimonious about including covariates in the model and does not exhibit the property of the AIC (or AICc) where more

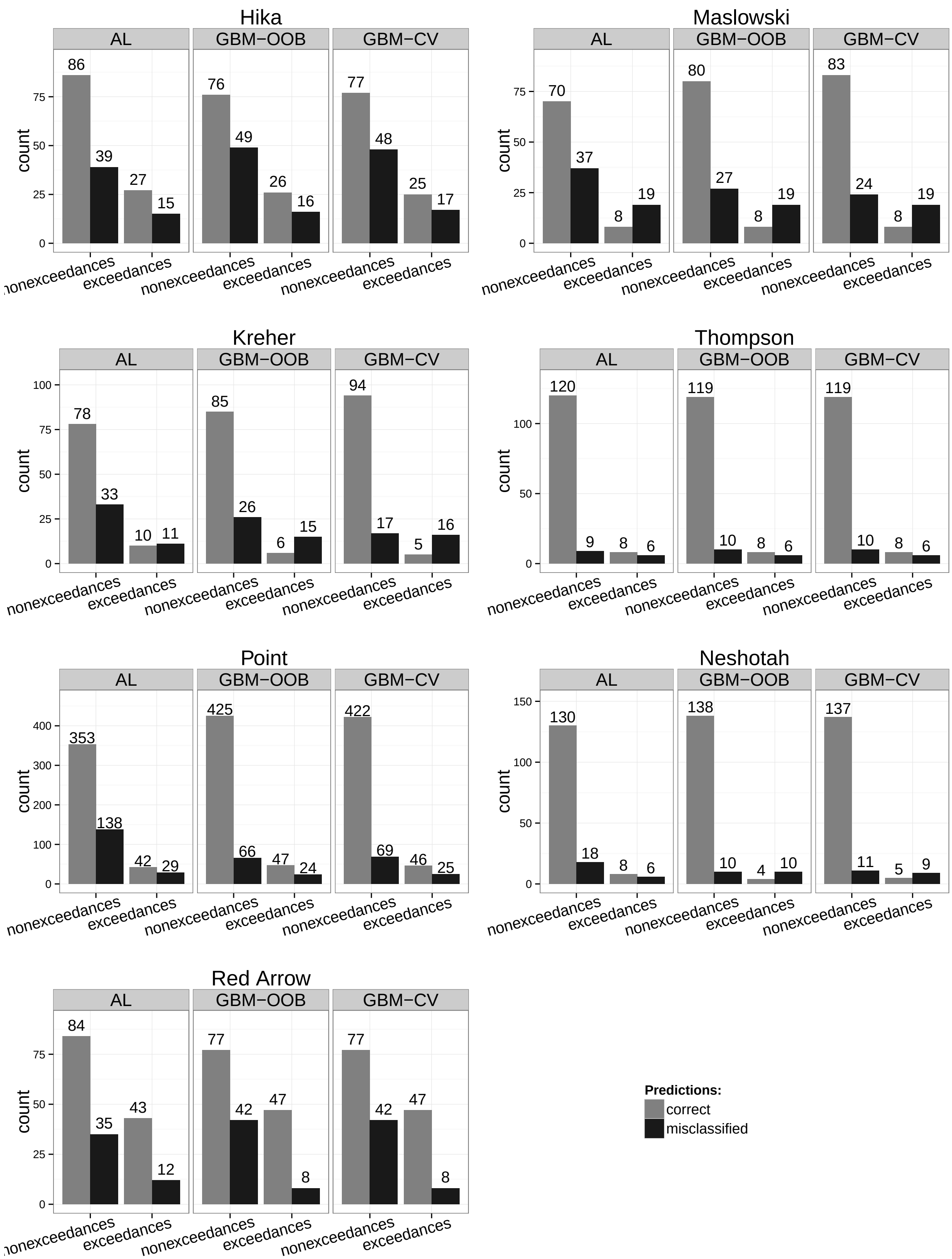

Figure 3: At each site, the number of predictions from AL, GBM-OOB, and GBM-CV that fell into four categories, from left: true negatives, false positives, true positive, and false negatives.

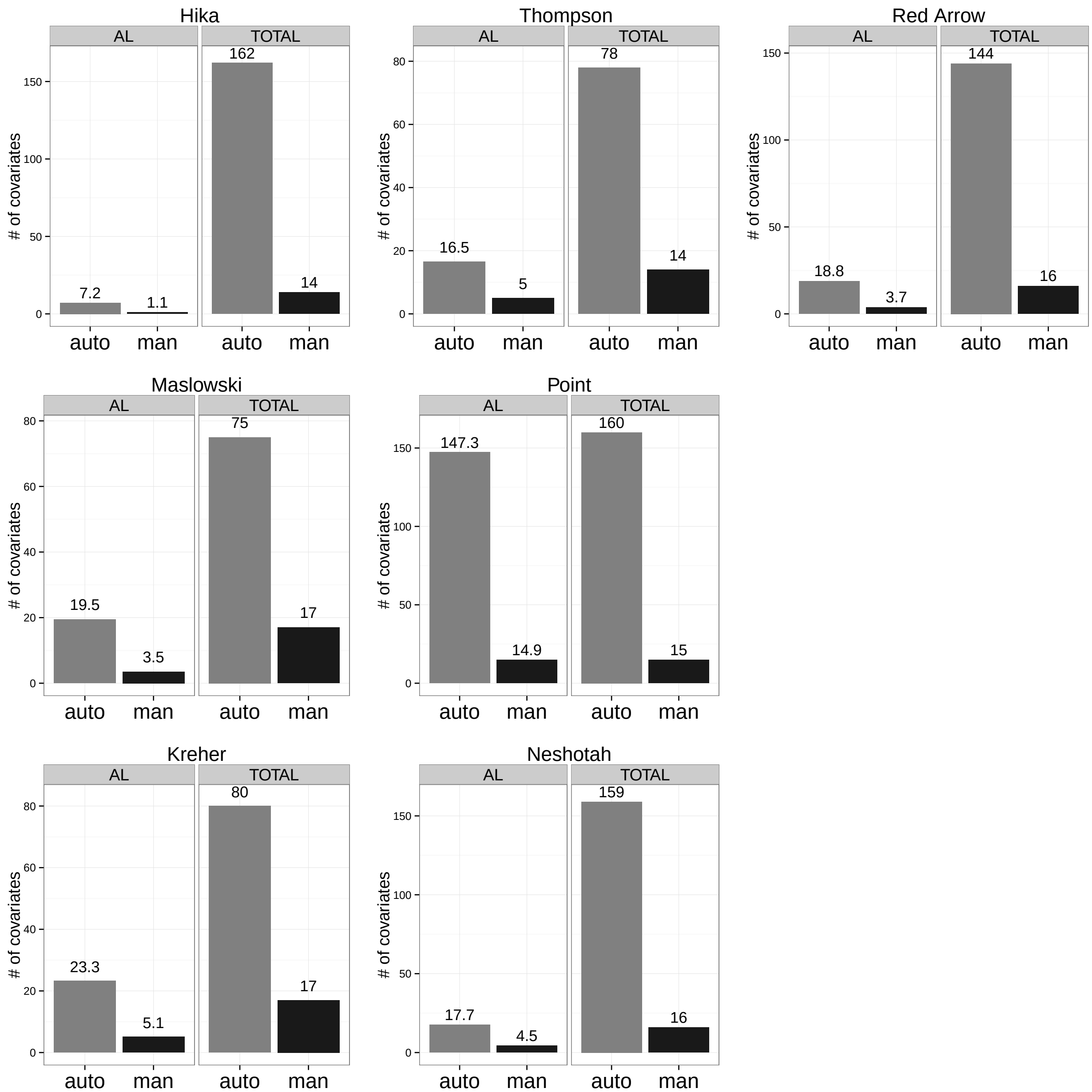

Figure 4: At each site, the mean number of covariates that were selected for the AL model, and the total number of covariates, all of which were used in the gradient boosting machine with an out-of-bag estimate of the optimal tree count (GBM-OOB) models. For both AL and GBM-OOB, the covariate counts are broken down by whether the covariate values were collected automatically from web services or manually at the beach.

covariates are included in the model as the number of observations increases. However, the BIC is derived from the standpoint of identifying the most probable model, rather than minimizing the prediction error. It is therefore likely that an AL model using the BIC for variable selection will have slightly worse predictive performance than one using the AICc.

Another advantage of the AL over GBM-type models is interpretability. As a linear regression technique, fitting an AL model means generating a set of coefficients, which can be interpreted as the marginal effect of a change in the corresponding covariate. On the other hand, GBM produces black-box models that typically make more accurate predictions but are difficult to intrepret. One common way to interpret a random forest model (such as from the GBM algorithm) is to observe the proportion of splits in the underlying trees that involve a particular covariate. The split proportion is a measurement of that covariate's importance to the model but gives no indication of how that covariate affects the bacterial concentraion.

All statistical methods and the comparison for this study were carried out in the R statistical software environment (R Core Team 2014). Scripts and details of the how the modeling methods were implemented are in the online supplement. Often times, beach management practitioners are not very familiar with statistical analysis and rely on more accessible software to help guide them through development of models for recreational water quality predictions. For this purpose, the Virtual Beach software was developed [Cyterski et al. (2013). Through version 2.4, the only method available in the Virtual Beach software was GA. As of version 3.0, Virtual Beach includes implementations of GBM, GA, and PLS models for prediction of bacterial concentration. An implementation of AL is also an anticipated addition to Virtual Beach.

7 Acknowledgments

The predictive models for this study were generated on facilities and software (HTCondor) provided by the University of Wisconsin-Madison's Center for High Throughput Computing.

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.

Brandt, S., D. Schwab, T. Croley, D. Belestky, and R. Whitman. *Ecosystem Forecasting: Integrating Science to Reduce the Risks to Human Health*. Vol. 87. 36; Suppl. American Geophysical Union.

Brauwere, Anouk de, Nouho Koffi Ouattara, and Pierre Servais. 2014. "Modeling Fecal Indicator Bacteria Concentrations in Natural Surface Waters: a Review." *Critical Reviews in Environmental Science and Technology*. http://www.tandfonline.com/doi/pdf/10.1080/10643389.2013.829978.

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.

Cabelli, V. J. "Health Effects Criteria for Marine Recreational Waters." Research and Development EPA-600/1-80-031. Vols. EPA-600/1-80-031. U.S. Environmental Protection Agency.

Cabelli, V. J., A. P. Dufour, M. A. Levin, L. J. McCabe, P. W. Haberman, and L. D. Jensen. "Relationship of Microbial Indicators to Health Effects at Marine Bathing Beaches." *Am.J.Public Health, 69(7), 690-696.*

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

"Colilert Test Kit Procedure." National environmental methods index. https://www.nemi.gov/methods/method_summary/5584/.

Cyterski, Mike, Wesley Brooks, Mike Galvin, Kurt Wolfe, Rebecca Carvin, Tonia Roddick, Mike Fienen, and Steve Corsi. 2013. *Virtual Beach 3: User's Guide*. United States Environmental Protection Agency. http://www2.epa.gov/exposure-assessment-models/virtual-beach-vb.

Dufour, A. P. "Health Effects Criteria for Fresh Recreational Waters." Vols. EPA-600/1-84-004.

Fleisher, J. M., L. E. Fleming, H. M. Solo-Gabriele, J. K. Kish, C. D. Sinigalliano, L. Plano, S. M. Elmir, et al. "The BEACHES Study: health Effects and Exposures from Non-Point Source Microbial Contaminants in Subtropical Recreational Marine Waters." *International Journal of Epidemiology*, 8–8.

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

Francy, Donna S., and Robert A. Darner. "Nowcasting Beach Advisories at Ohio Lake Erie Beaches." *Open-File Report*. Vols. Open-File Report 2007-1427. U.S. Geological Survey.

Francy, Donna S., Amie M. G. Brady, Rebecca B. Carvin, Steven R. Corsi, Lori M. Fuller, John H. Harrison, Brett A. Hayhurst, et al. 2013. "Developing and Implementing the Use of Predictive Models for Estimating Water Quality at Great Lakes Beaches." USGS Scientific Investigations Report 2013-5166.

Frick, W. E., Z. Ge, and R. G. Zepp. "Nowcasting and Forecasting Concentrations of Biological Contaminants at Beaches: A Feasibility and Case Study." *Environ.Sci.Technol.* 42 (13): 4818–24.

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

Ge, Z., and W. E. Frick. "Some Statistical Issues Related to Multiple Linear Regression Modeling of Beach Bacteria Concentrations." *Environ.Res.* 103 (3): 358–64.

Hanley, James A., and Barbara J. McNeil. 1982. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve." *Radiology* 148 (1): 29–36.

He, Li-Ming Lee, and Zhen-Li He. "Water Quality Prediction of Marine Recreational Beaches Receiving Watershed Baseflow and Stormwater Runoff in Southern California, USA." *Water Research* 42 (10-11): 2563–73.

Hosmer Jr, David W, and Stanley Lemeshow. 2004. *Applied Logistic Regression*. John Wiley & Sons.

Hou, D., S. J. M. Rabinovici, and A. B. Boehm. "Enterococci Predictions from Partial Least Squares Regression Models in Conjunction with a Single-Sample Standard Improve the Efficacy of Beach Management Advisories." *Environ. Sci. Technol.* 40 (6): 1737–43.

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.

Jin, G., and A. J. Englande Jr. "Prediction of Swimmability in a Brackish Water Body." *Manage.Environ.Qual.* 17 (2): 197–208.

Jones, Rachael M., Li Liu, and Samuel Dorovitch. 2012. "Hydrometeorological Variables Predict Fecal Indicator Bacteria Densities in Freshwater: data-Driven Methods for Variable Selection." *Environmental Monitoring and Assessment* 185: 2355–66. http://link.springer.com/article/10.1007/s10661-012-2716-8.

Kashefipour, S. M., B. Lin, and R. A. Falconer. 2005. "Neural Networks for Predicting Seawater Bacterial Levels." *Proceedings of the Institution of Civil Engineers-Water Management* 158: 111–18. http://www.icevirtuallibrary.com/content/article/10.1680/wama.2005.158.3.111.

Nevers, M. B., and R. L. Whitman. "Nowcast Modeling of Escherichia Coli Concentrations at Multiple Urban Beaches of Southern Lake Michigan." *Water Research* 39 (20): 5250–60.

Olyphant, G. A., and R. L. Whitman. "Elements of a Predictive Model for Determining Beach Closures on a Real Time Basis: the Case of 63rd Street Beach Chicago." *Environ.Monit.Assess.* 98 (1-3): 175–90.

Parkhurst, D. F., K. P. Brenner, A. P. Dufour, and L. J. Wymer. "Indicator Bacteria at Five Swimming Beaches - Analysis Using Random Forests." *Water Res.* 39 (7): 1354–60.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Richard Zepp, Michael Cyterski & Shuyan Zhang & Emily White & Marirosa Molina & Kurt Wolfe & Rajbir Parmar &. 2012. “Temporal Synchronization Analysis for Improving Regression Modeling of Fecal Indicator Bacteria Levels.” *Water, Air &Soil Pollution* 223: 4841–51.

Stidson, R. T, C. A. Gray, and C. D. McPhail. 2012. “Development and Use of Modelling Techniques for Real-Time Bathing Water Quality Predictions.” *Water and Environment Journal* 26: 1747–6593. http://onlinelibrary.wiley.com/doi/10.1111/j.1747-6593.2011.00258.x/abstract;jsessionid=BD291BB8B49C5EBBA2F8CABD79A23F2E.f03t02.

Thoe, W., M. Gold, A Griesbach, M. Grimmer, M. L. Taggart, and A. B. Boehm. 2014. “Predicting Water Quality at Santa Monica Beach: evaluation of Five Different Models Fror Public Notification of Unsafe Swimming Confditions.” *Water Research* 67: 105–17.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society, Series B (Methodological)*, 267–88.

U.S. Geological Survey. 2014. “Environmental Data Discovery and Transformation.” September. http://cida.usgs.gov/enddat/.

USEPA. 1986. “Ambient Water Quality Criteria for Bacteria.” EPA440/5-84-00.

———. 2008. “Great Lakes Beach Sanitary Survey User Manual.” Washington, DC: US Environmental Protection Agency, Office of Water. http://www2.epa.gov/sites/production/files/documents/greatlakesmanual.pdf.

———. 2012. “Recreational Water Quality Criteria.” EPA-820-F-12-058. USEPA. http://water.epa.gov/scitech/swguidance/standards/criteria/health/recreation/upload/RWQC2012.pdf.

———. “Critical Path Science Plan for the Development of New or Revised Recreational Water Quality Criteria.”

Wade, T. J., R. L. Calderon, E. Sams, M. Beach, K. P. Brenner, A. H. Williams, and A. P. Dufour. “Rapidly Measured Indicators of Recreational Water Quality Are Predictive of Swimming-Associated Gastrointestinal Illness.” *Environmental Health Perspectives* 114 (1): 24–28.

Wade, Timothy J., Rebecca L. Calderon, Kristen P. Brenner, Elizabeth Sams, Michael Beach, Richard Haugland, Larry Wymer, and Alfred P. Dufour. “High Sensitivity of Children to Swimming-Associated Gastrointestinal Illness: results Using a Rapid Assay of Recreational Water Quality.” *Epidemiology (Cambridge, Mass.)* 19 (3): 375–83.

Waschbusch, R., S. Corsi, K. Sorsa, J. Walker, J. Standridge, and T. Schnieder. “Data Collection and Modeling of Enteric Pathogens, Fecal Indicators and Real-Time Environmental Data at Madison, Wisconsin Recreational Beaches for Timely Public Access to Water Quality Information.” *Stormwater: The Journal for Surface Water Quality Professionals* Final report for the EMPACT Project R-82933901-0.

WDNR. 2012. WDNR. http://dnr.wi.gov/topic/Beaches/documents/BeachReport2012.pdf.

Whitman, R. L., and M. B. Nevers. “Summer E. Coli Patterns and Responses Along 23 Chicago Beaches.” *ENVIRONMENTAL SCIENCE & TECHNOLOGY* 42 (24): 9217–24.

———. “Escherichia Coli Sampling Reliability at a Frequently Closed Chicago Beach: Monitoring and Management Implications.” *Environ.Sci.Technol.* 38 (16): 4241–46.

Whitman, R. L., M. B. Nevers, G. C. Korinek, and M. N. Byappanahalli. “Solar and Temporal Effects on Escherichia Coli Concentration at a Lake Michigan Swimming Beach.” *Appl.Environ.Microbiol.* 70 (7): 4276–85.

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.