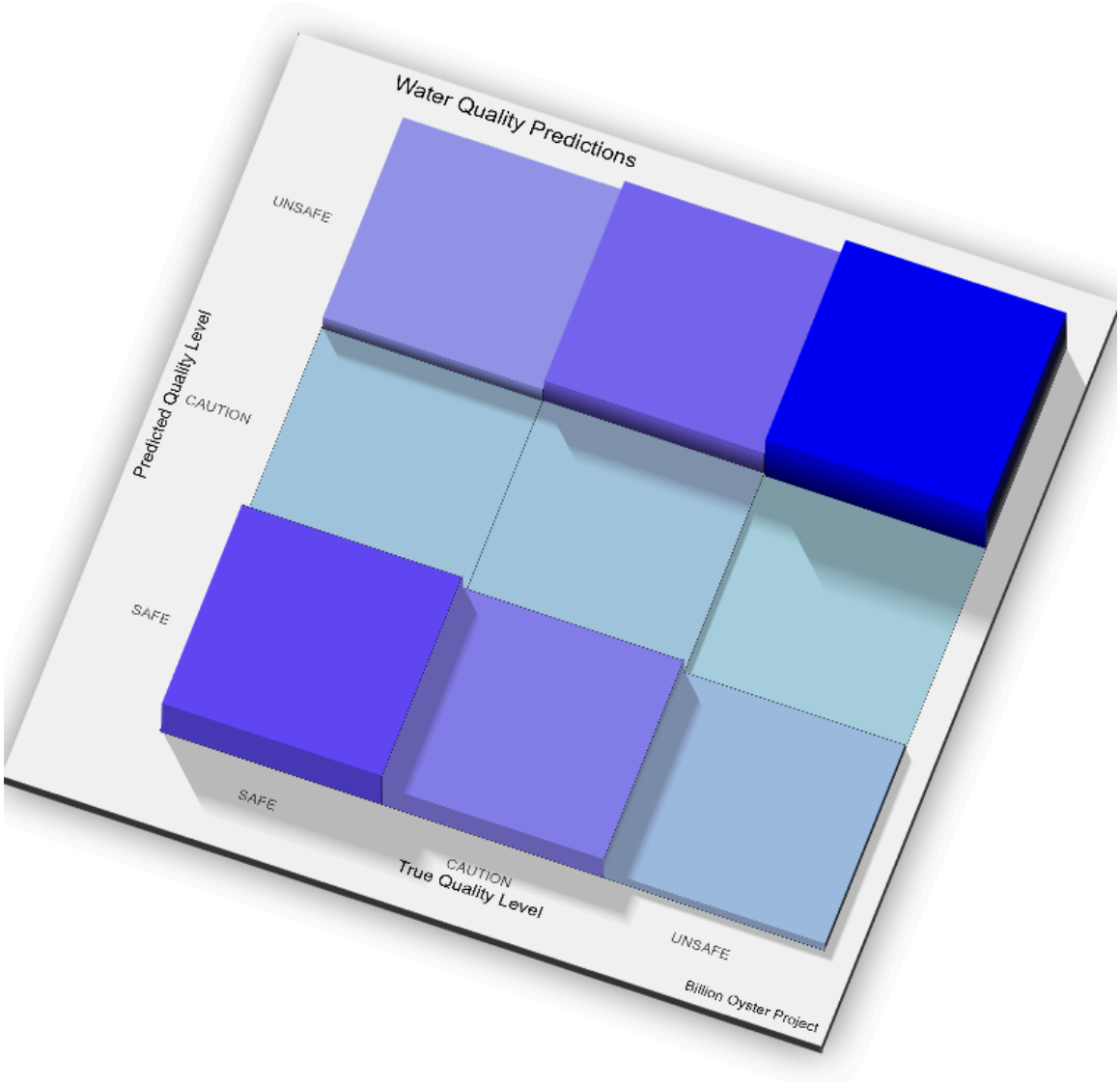
Predicting Enterococci Levels in New York Harbor

A Study for the Billion Oyster Project by Arthur Steinmetz

2024-08-07



Abstract

This document explores the relationship between testing location, weather, tides and water quality in the NYC Harbor. The data sources are the Billion Oyster Project (BOP), the Citizens' Water Quality Testing Program and the NOAA. The data spans from Autumn 2011 to June 27, 2024. The stations reporting vary over time.

This note is an update of a previous exploration of the data where I showed that a linear regression model was a very poor explainer of bacteria levels in New York Harbor water. In this note I set a lower bar by classifying bacteria into just three classes, "safe," (<=35 colonies) "caution" (<= 104 colonies) and "unsafe" (> 104 colonies). I train a "random forest" machine learning model on a sub-sample of the data and then evaluate the model with a different test set of data.

In summary, this model works very well in fitting the training set but does much worse out of sample. The model does show good accuracy in predicting "safe" and "unsafe" water but very little accuracy in predicting bacteria levels in the "caution" range. The dominant predictor is the testing site, since several sites NEVER have "safe" water in the data set. No other variable stands out in significance.

This is not an academic-quality study. It is an exploration of the data. I am not a water quality expert or a professional statistician. Comments and criticism are welcome.

Data

The main data source is the BOP water quality spreadsheet found here: BOP Water Quality Data[1] I also used the NOAA data site for tide, temperature and rainfall data.

Feature Engineering

The BOP data includes time of last high tide. I thought I could get more granular by imputing the direction and strength of the tidal current at the time of the water sample. I used the NOAA tide data from the nearest statoin to find the previous slack tide time and level, then the next slack tide time and level. By determining where in the tide phase the sample was taken and the total change in water level for that phase, I impute the direction and strength of the tidal current when the sample was taken using this formula:

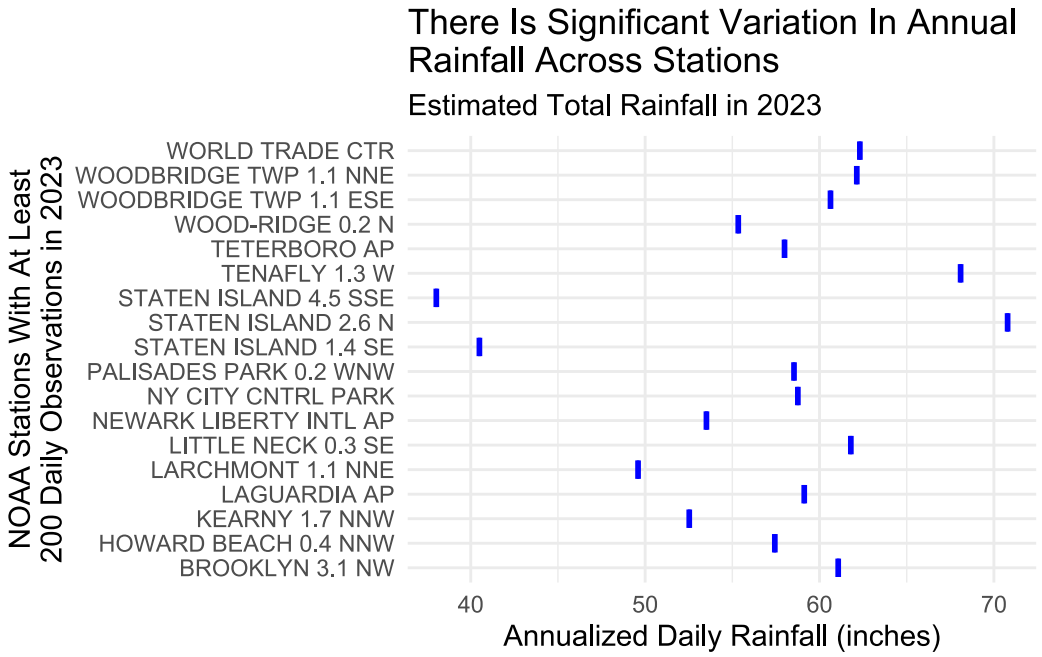
$$CurrentSpeed = HighLowRangeFt * sin\left(\pi * \frac{HoursSinceLastTide}{TideDurationHrs}\right)$$

So the further we are from a slack tide, high or low, the faster the current will be. The bigger the change in water level during a tidal phase, the stronger the current will be. Ebb tides are negative values, flood tides are positive. *CurrentSpeed* is an index so the units don't have a specific meaning like feet-per-second.

The city of New York uses 48-hour rainfall amounts in its safety criteria so that is what I use as one precipitation variable. Since the sample time is in the middle of the day, that day's rainfall

[1] https://docs.google.com/spreadsheets/d/1813b2nagaxZ80xRfyMZNNKySZOitro5Nt7W4E9WNQDA/edit?gid=1924583806#gid=1924583806

might not be relevant. I choose the prior two days for the 48-hour period. I also use the current day rainfall so there are three days of precipitation in the dataset. I use the NOAA rainfall data for the closest station to the sampling site. Where the location of the sampling site is not known, I default to the Central Park weather station. Station location is more important than it might first appear. Rainfall varies widely by location in the city.



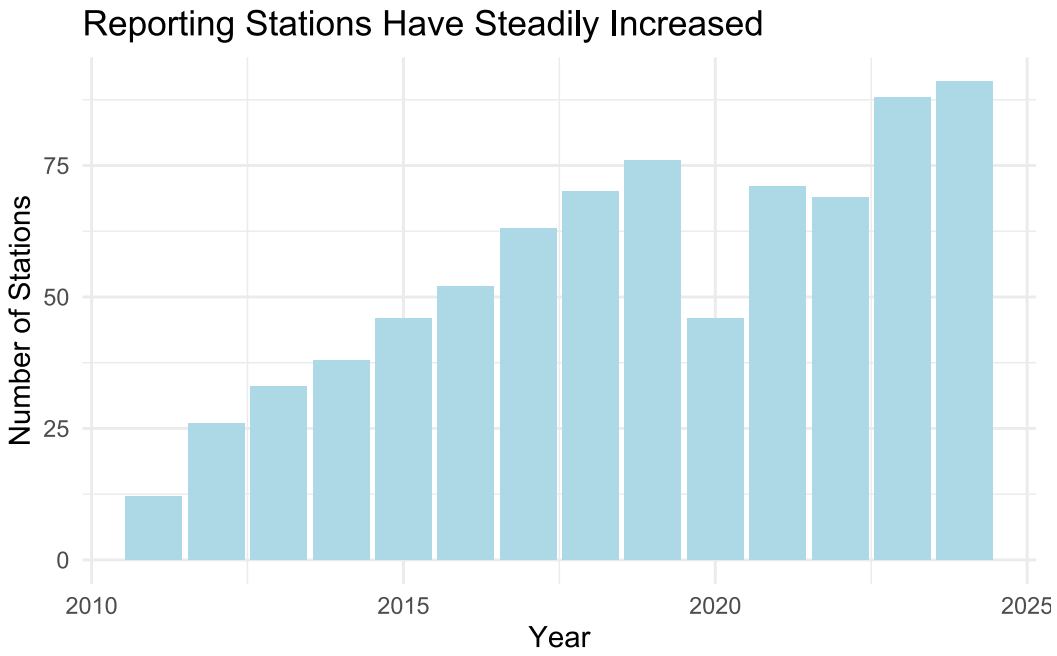
The BOP data does not include temperature. I used the NOAA Air temperature at the nearest station for each sample day as a data feature. This is a (not very good) proxy for the water temperature but also for seasonality. This allows seasonality to be a continuous variable. Otherwise, "month" would be a categorical variable but we can omit it.

The NOAA Weather and tide data I retrieved are substituted for the equivalent features in the BOP water quality spreadsheet. This is because many of the BOP observations default to either Central Park or Battery Park, irrespective of the closest NOAA station. This leaves us with the following data features:

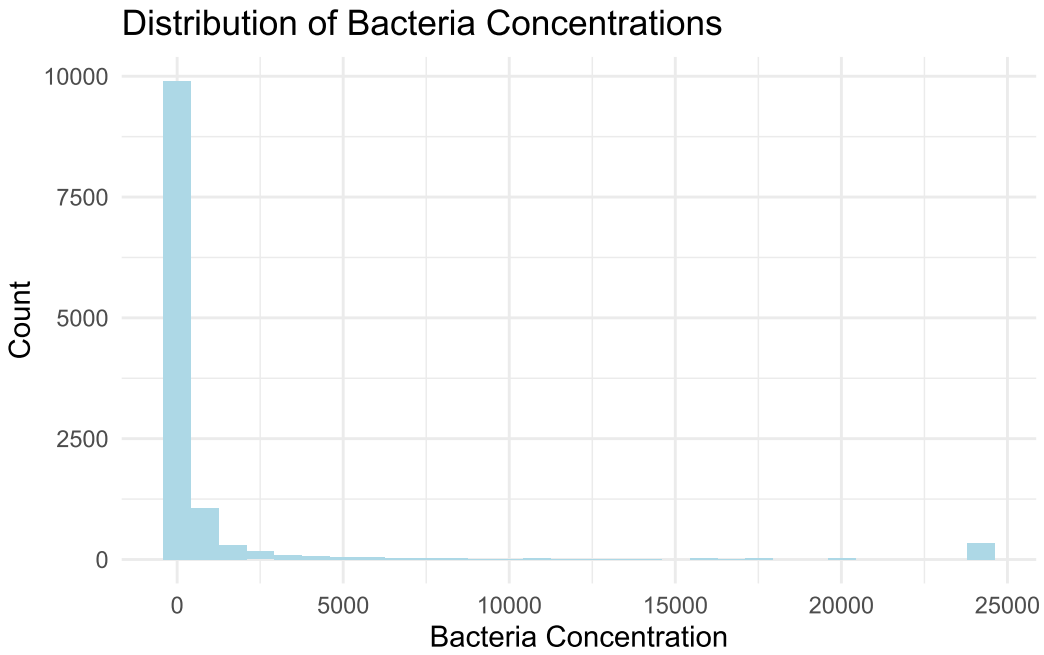
```
 [1] "bacteria"                          "quality"
 [3] "site_id"                           "site"
 [5] "date"                              "water_body"
 [7] "neighborhood_town"                 "nys_dec_water_body_classification"
 [9] "nyc_dep_wrrf_or_sewershed"         "tide_level"
[11] "hours_since_last"                  "current"
[13] "temperature_f"                     "ghcn_precip_in"
[15] "ghcn_precip_in_48"
```

Data Exploration

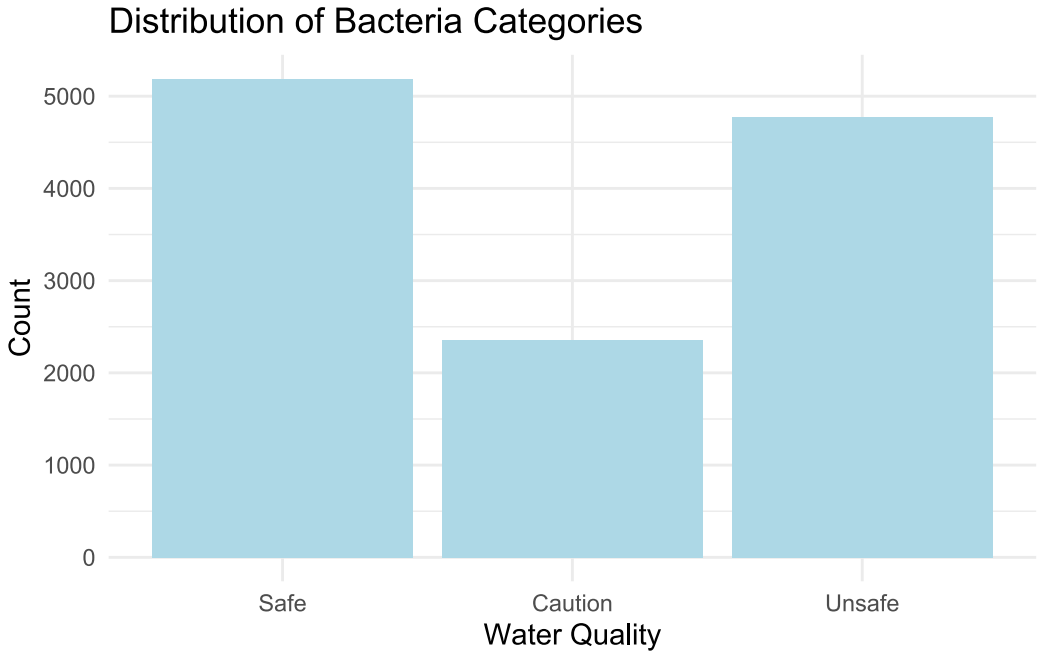
As mentioned above, the number of stations reporting varies over time. It has grown steadily though during the COVID crisis fewer stations reported.



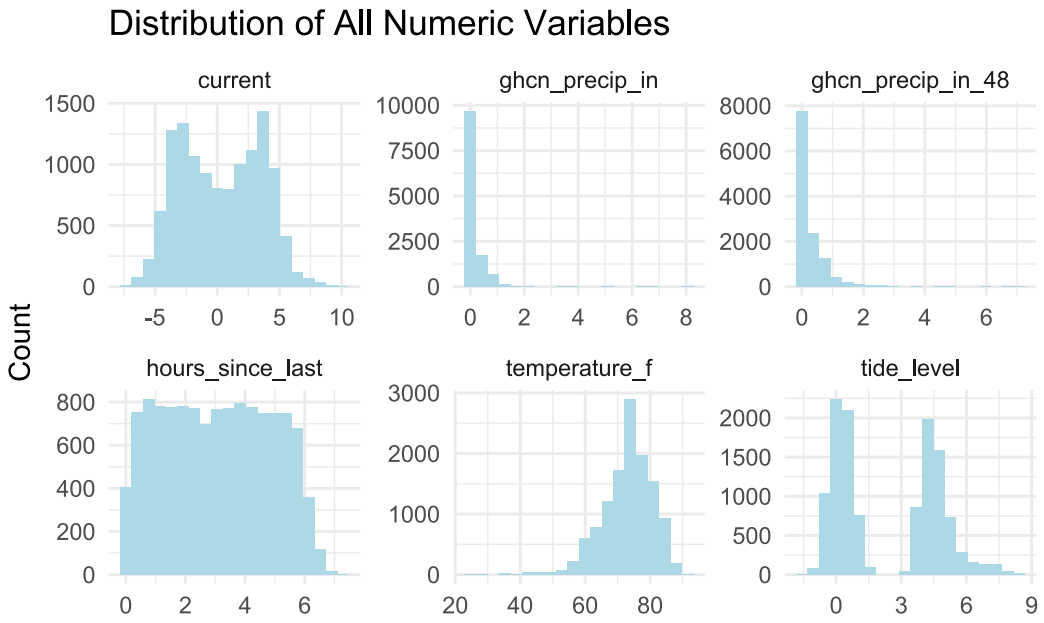
The bacteria levels are distributed in a lopsided way. The extreme high level is effectively infinity and conveys little information. Values above 5000 are only 5% of the observations and values below 500 are 82% of the observations.



If we group the bacteria levels according to the official quality standards we get a better behaved distribution. As a result, we will attempt to predict the quality classification of the water rather than the bacteria level.

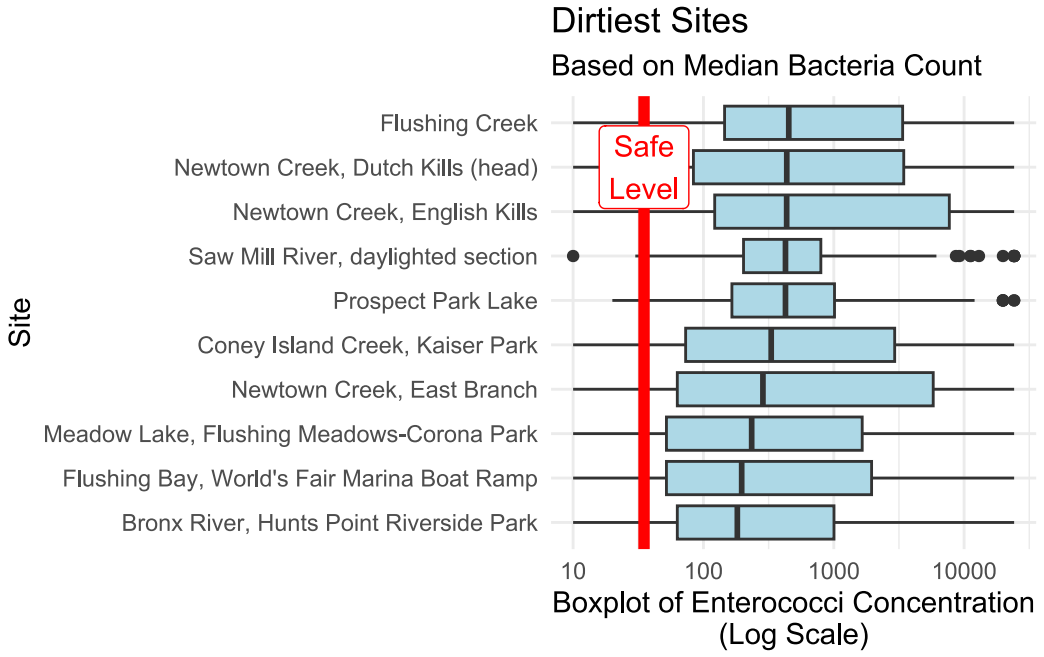


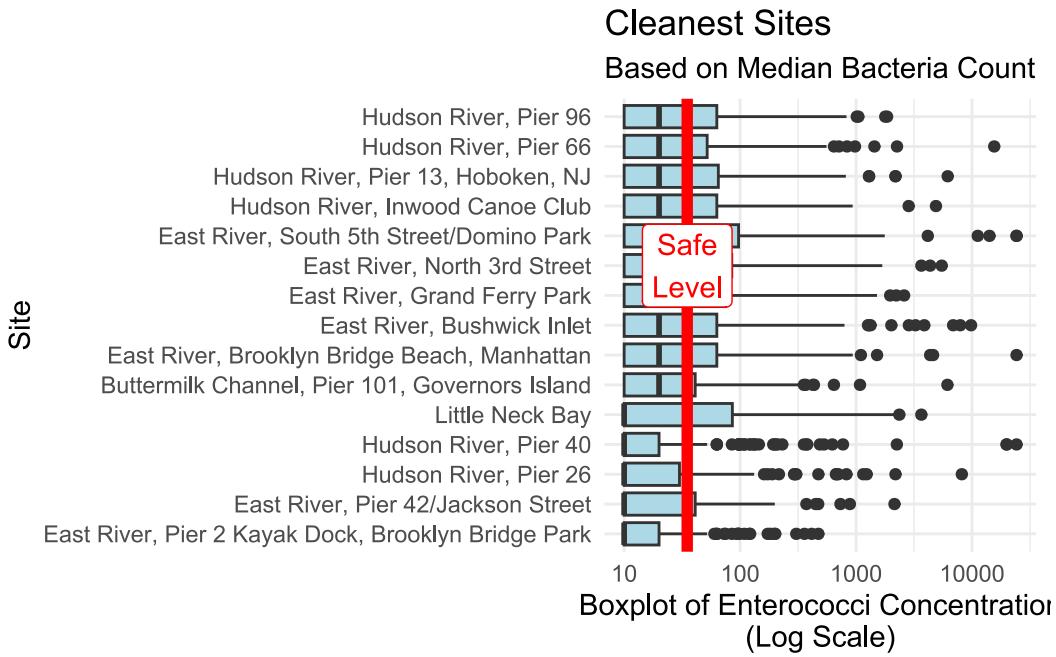
What are the distributions of all the variables? Note the tide level distribution shows the levels at just the high and low tides but we know when in the tidal phase the sample was taken.



The remaining features are categorical, not continuous variables, so we will have to take that into account when we build our model.

What are the cleanest and most contaminated sites? The boxes in the plots show the median and inter-quartile ranges.

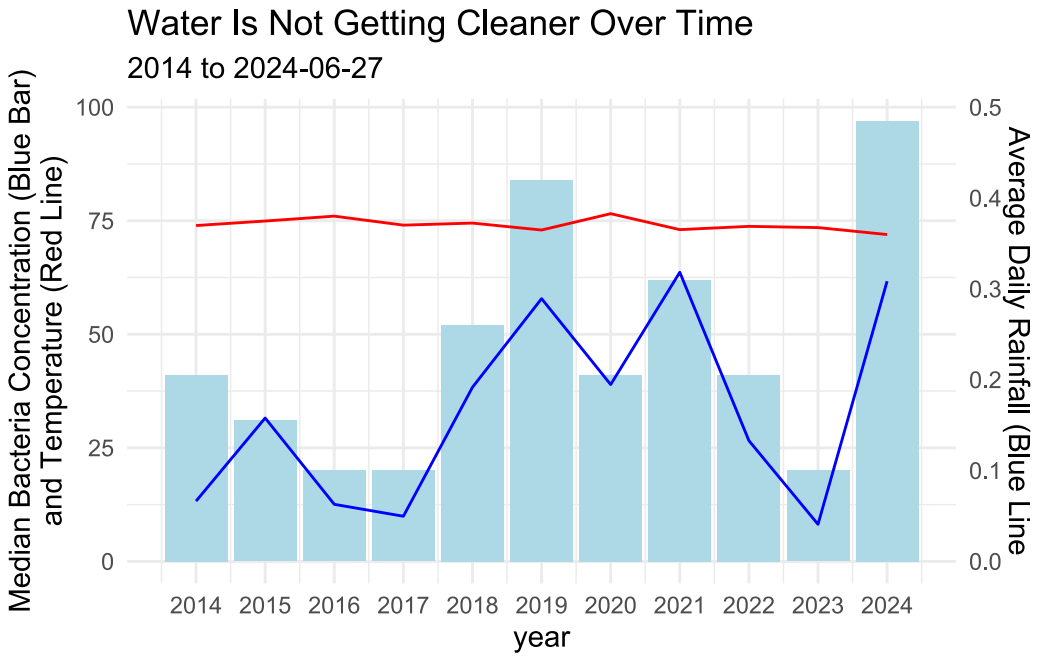




What is obvious is that even the cleanest sites have a lot of variation in bacteria levels. This might give us some hope that environmental factors might be more important than location in predicting bacteria levels.

Now let's look at some trends over time. Sadly, the overall level of bacteria has not improved over time. Looking at temperature, there are no clear trends. There does seem to be a relationship between rainfall and bacteria. Note the rainfall measurements are from the closest NOAA site on the day of the water sample, so it may not be a proxy for annual rainfall.

As we saw, the number of reporting stations has increased so the data may not be comparable year-on-year. Let's look at just the stations that have been reporting for at least 10 years.



Modeling

We use a random forest algorithm to train a prediction model. This class of models works very well on imbalanced data like we have here. It can also handle data sets with many categorical inputs like `site` in this case. More on this technique can be found at https://en.wikipedia.org/wiki/Random_forest . To create the model we split the data randomly into a training set and a test set. 75% is used for training and the rest we hold out for testing. The sets are stratified so the same proportion of each bacteria quality is in each set. The model is tuned using cross-validation on the training set and then evaluated on the test set.

All of the code and data used to create the model can be found in the project's GitHub repository at https://github.com/apsteinmetz/oyster.git .

Results

The simplest way to evaluate the model is to look at the confusion matrix. This is a table that shows the number of correct and incorrect predictions for each quality. In a perfect model all the observations would lie on the diagonal and the off-diagonal counts would all be zero. The table below shows that out of 1194 "UNSAFE" observations, the model predicted or 76%, correctly. This was the best result. While overall accuracy is important, we might be most concerned about cases when the model predicts "SAFE" water when it's not. In 8% of *all* the cases, the model predicted the water was "SAFE" when the actual was "UNSAFE" (239/3082). Additionally, The model is far better at predicting "SAFE" and "UNSAFE" than "CAUTION."

Truth Table

Truth	Prediction			
	Safe	Caution	Unsafe	Total
Safe	75%	4%	21%	100%
Caution	53%	8%	39%	100%
Unsafe	20%	4%	76%	100%

Proportions

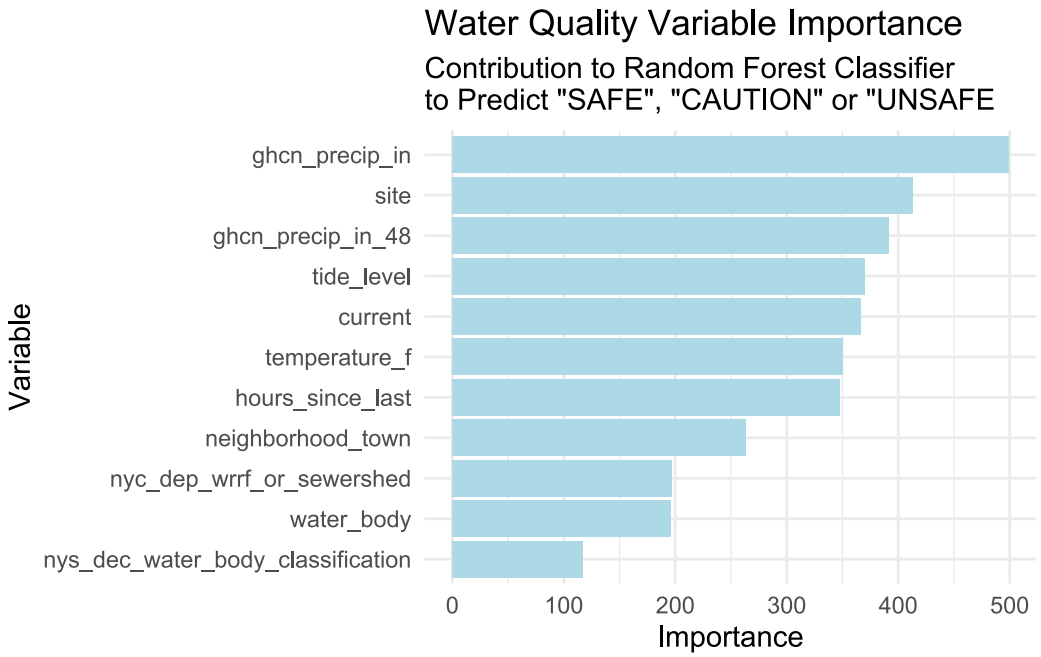
Truth Table

Truth	Prediction			
	Safe	Caution	Unsafe	Total
Safe	977	48	273	1,298
Caution	314	46	230	590
Unsafe	239	42	913	1,194
Total	1530	136	1416	3082

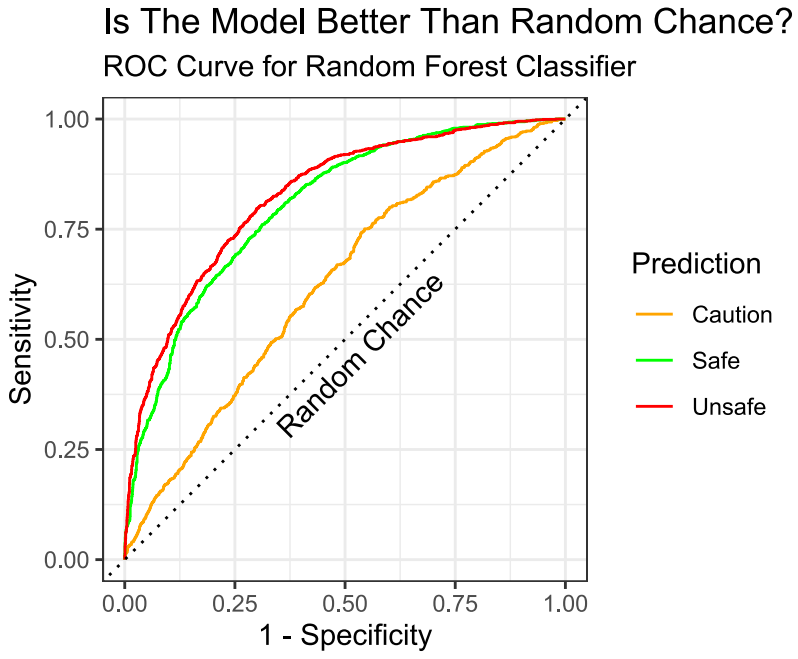
Counts

In a regression analysis, we measure the coefficient of each of the input variables so there is a precise measure of how much each variable contributes to the prediction. In a random forest, we don't have linear relationships but we still can measure the relative importance of each variable.

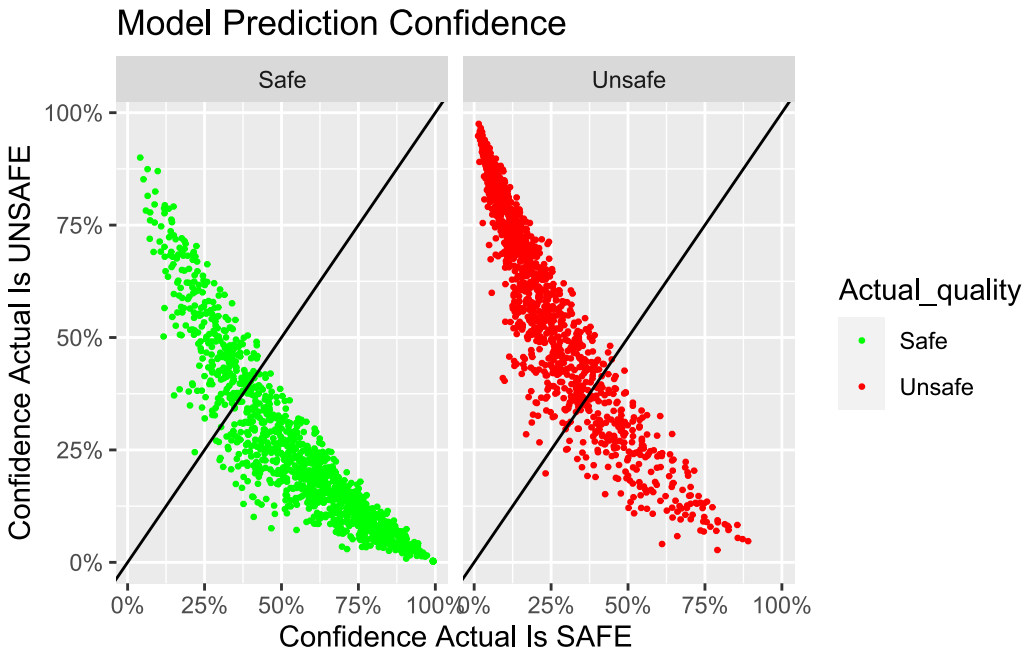
Daily rainfall is the most important determinant of water quality. This is not surprising since heavy rainfall overwhelms the sewage system and washes bacteria into the water. The `site` input is the second most important feature. Since location doesn't change over time, we don't need a model to tell us some spots are worse than others. Somewhat surprisingly, then, `sewershed` is not particularly important.



The "Receiver Operator Curve" visualizes how much better than model is than random chance. Curves that bend up and to the left are better. We can see that the "Caution" predictions are barely better than a coin flip. Overall results can be summarized with the the single *"Kappa"* statistic which indicates the model is about 39 percentage points better than random chance.



Every prediction the model makes includes a confidence level. The quality with the highest confidence level is what the model predicts. How certain is the model that the prediction is correct? The plot below shows the confidence level for each prediction. Each panel contains a dot for each observation, divided into the actual classifications. The position of each dot shows the relative confidence the model has that the water is "Safe" and "Unsafe." The observations and confidence values for "Caution" are not shown. The majority of the quality classifications are correct but we can see the model is highly confident in many cases where it is wrong.



Conclusion

We have created a random forest model that shows modest accuracy in predicting whether enterococci levels will be at the extremes of "safe" or "unsafe." The model is not very good at predicting the middle "caution" levels of bacteria. We have shown that rainfall is the most important determinant of water quality. Unfortunately, the site location itself is an extremely important predictor of bacteria levels so changing environmental factors like tide and temperature tell us little about levels over time. Understanding point sources of pollution around each site and how they vary over time seems like a good next step but the fact that `sewershed` is not a very important predictor is discouraging.

Note

We can illustrate the importance of separating training and testing data. It's easy to "overfit" when including all of the data. In the example below we train on all of the water quality data. The accuracy and prediction confidence are very high but it's an illusion. We don't know anything about predictive ability in the future.

Prediction is "Easy" When We Overfit

Truth Table

	Prediction			
Truth	Safe	Caution	Unsafe	Total
Safe	97%	0%	3%	100%
Caution	15%	76%	10%	100%
Unsafe	4%	0%	96%	100%