# GAM and random forest models of dogfish bottom trawl catches

offered for SSC Groundfish Subcommittee Sep 29-30 meeting

Corey Niles WDFW

9/28/2021

## Overview

This report presents modelling results relevant to the question of seasonality and the availability of dogfish to bottom trawl gear. Two flexible model types are used to explore the non-linear effects of area: generalized additive models (GAMs) and regression trees.

I had not planned on performing this analysis at this time or submitting it to the SSC's Groundfish Subcommittee. After receiving the STAT's report summarizing their generalized linear model last Thursday and reviewing their model and conclusions, I felt that the SSC's discussion could benefit from what's presented here. Their conclusions appear contrary to patterns I was seeing in modeling dogfish catches for purposes of 2023-2024 management measure options.

This is an abbreviated write-up and time for review is short. The pattern I wish to highlight is shown in Figure 1 below. In sum, estimated catch rates are multiple times higher in Nov-Jan than in other months of the year and lowest in the period when the survey is taking place. Fishery data is noisy of course. Yet the reasons why catch would be so notably higher other than increased density/availability of the fish in the area are not clear.

In addition to different model types, this analysis also differs from the STAT's in that it uses hurdle models to separately model presence/absence and size of the catch when encountered. And beyond just considering the models for the fishery data, I apply the models to the locations of the NWSFC bottom trawl survey to explore how predicted catch changes under the counterfactual situation that each tow happened in each month of the year.

I admit to not understanding how the SSC would use this information to construct a prior on q. And I'm not intending to argue for a particular result. I offer the results believing they would make for a fuller discussion.

All in all, we appreciate the efforts of the SSC and STAT to further evaluate the issue. And in the bigger picture, I fully agree with the STAT's conclusion, expressed to the Subcommittee in August, that transboundary work will be needed to resolve the issue more fully. WDFW has long supported transboundary science and assessment for dogfish and many other stocks and recognizes the challenges. As for this particular analysis, with limited time, I did not have opportunity to provide our SSC, GMT, or other PFMC team members time for feedback and comment.

## Methods

As time is limited, this is only a quick sketch of the methods. I performed all analyses in R and would happily share the code. Data access is granted by WCGOP. I provide some of the model input at the end of this document.

## The data

The dataset is the same used by the STAT and includes 2002-2019 WCGOP observations of bottom trawl tows/hauls. I filtered for all bottom trawl gears in the Limited Entry Trawl, Catch Shares, and Catch Shares EM sector filed (i.e. using the "gear" and "sector" fields) and filtered out "Trip Without Catch" records. Dogfish catches of less than 0.22 pounds were changed to 0.22 pounds.[1]

In addition, to eliminate the influence of targeting, I excluded hauls from trips where any revenue was recorded on fish tickets. This involved matching the fish ticket numbers in the WCGOP data set to the PacFIN database. Only 2,478 of the 96,721 observed hauls are associated with trips where dogfish revenues were recorded. Dropping them is a conservative step as most revenues are small and likely not indicative of targeting or less of an incentive to avoid dogfish.

## Model approach

For the GAMs, I used the mgcv package. For the random forest models, I used the ranger package in conjunction with the tidymodels framework for evaluating the tuning parameters using the test, training, and validations sets.

For both model types, I use a hurdle model approach where the presence/absence of dogfish and the size of catch are modeled separately. For the size of catch, both models assume a lognormal distribution because of the highly skewed pattern created by low frequency but very large catches of dogfish.

The predictor variables include location of the set (in UTM 10N, WGS84 coordinates), set depth, haul duration, month, and year. The variables are kept the same between all models with the exception that month is modeled as a factor (and a non-smooth term) in the GAMs. The way the random forest tree-based method works, I left month as numeric. Also, the random forest classification model requires the presence/absence to be a factor variable whereas the GAMs work off 1s and 0s.

With the fitted models, I predict catch for each NWFSC survey tow successfully conducted 2003-2019. There are 10,870 tows in total with an average of 634.4 per year. The predicted catches are based on survey tow attributes except for duration and the counterfactual change of month. Duration is set to the median value from the WCGOP data. The model predicted catch is produced by multiplying the modeled probability of occurrence by the expected catch size when present.

# Results

Figure 1 shows the main results of interest. Again, the idea was to explore the question–what if each survey tow happened in a different month of year? What's the relative difference in expected catch compared to the month when the tow actually happened?

The values summarized on the y-axis are the ratio of predicted catch for each tow if it had occurred in the month on the x-axis to the predicted catch in the month the tow actually occurred. There are two distributions per month, one for each model type, and each consisting of 10,870 values for the months when the survey doesn't occur and between 8,147 and 9,848 for the months when it does. The distributions are summarized by the median, mean (where legible), and 25th-75th percentile intervals.

As can be seen, the GAM and random forest models differ in terms of scale but show the same relative pattern among the months. The models both estimate little difference in the months during which the survey occurs and large differences in the months where the survey does not occur or only partially occurs.

Figure 2 is offered to provide a sense of the spatial patterns. It shows the same distributions of ratios as in Figure displaying the smoothed conditional mean by latitude. Only Jan, Nov, and Dec are shown. As can

[1] This corresponds to ~100 grams, which corresponds to a 30 cm fish and the length at which the assessment estimated selectivity becomes non-zero for bottom trawl gear.

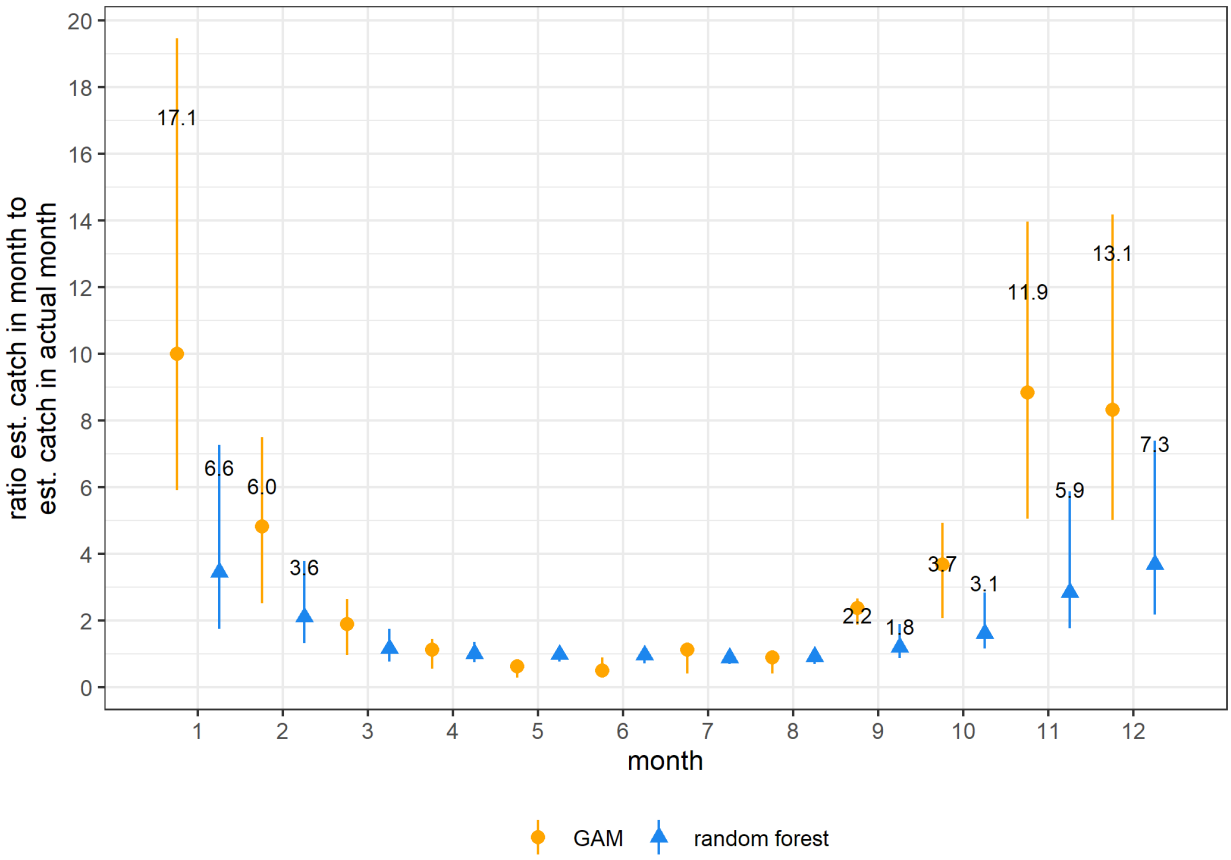



Figure 1: Ratio of model predicted catches at survey tow locations by month to predicted catch of the tow in the month it actually occurred. Median values are displayed (GAM = orange circles, random forest = blue triangles) with the lines indicating the central 50% of the distributions. The numeric values display the mean value (omitted for March-August because of illegibility).

be seen, the higher predicted catch appears across multiple areas of the coast. Both models see less of an increase in the northernmost latitudes than in other areas.

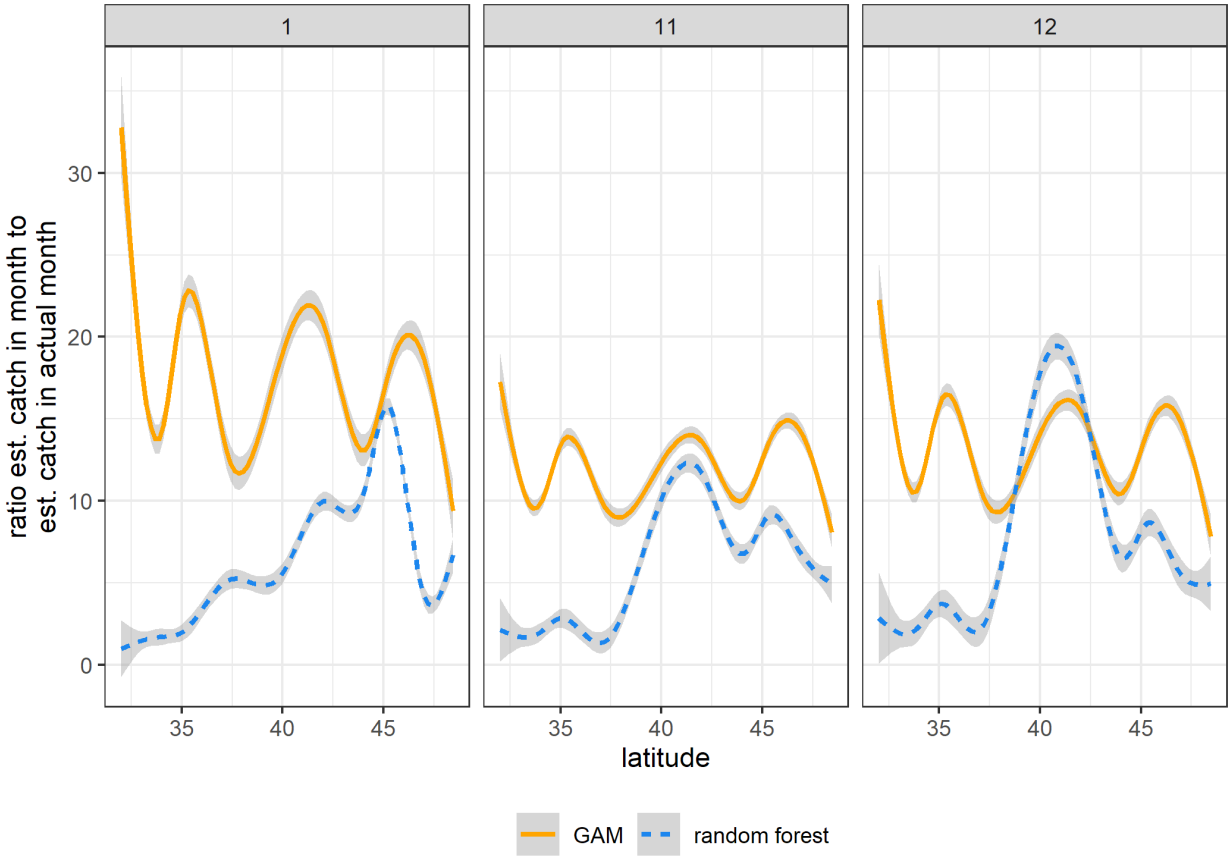



Figure 2: Smoothed mean values over latitude for the GAM (orange, solid line) and random forest (blue, dashed line) modeled ratios of predicted catch by month to actual month of the survey tow just for Nov-Jan. The curves are the default produced by ggplot2's geom_smooth, which is also an mgcv GAM

Figure 3 summarizes estimated catches by month for each model, applied to the original WCGOP data set, and compares them to the distribution of actual recorded WCGOP catch. Unlike the figures above, for this comparison the probability of occurrence in the hurdle models was rounded to better match the actual data (i.e. a value of 0.5 and above equals a catch event and below a zero). As can be seen, the modeled mean catch underestimate the mean catch by month in all cases. This illustrates the challenge of capturing the rare, very large hauls with a single distribution.

Lastly, Figure 4 plots the same distributions as in Figure 3 but as summarized by the smoothed conditional means for each month (each facet) and latitude. The pattern that sticks out is that the raw data estimates appear to be highly influenced by the northern latitudes in every single month. Comparing these to distributions shown in Figure 2 might explain some of the difference between the raw data and the counterfactual exercise of the model predictions by month the survey tow locations. Catch at the survey locations are modeled at their actual location and so the effect of month on the northern stations is limited to the northern stations.

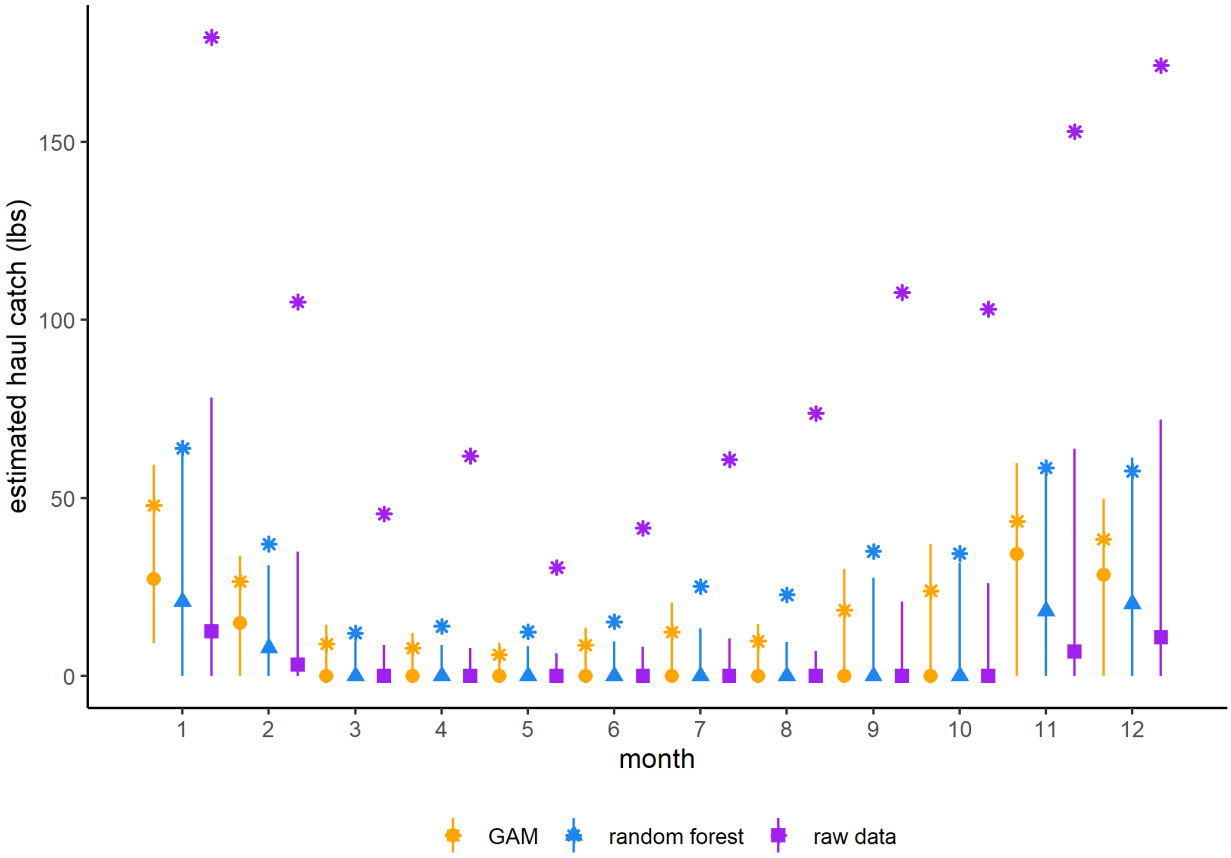



Figure 3: Comparison of monthly distributions of the actual data and GAM and random forest estimated catches for WCGOP observed bottom trawl hauls. The lines display the 25th-75th percentile intervals and the asterisk above each displays the mean. The median of each monthly distribution is displayed by the orange circle for the GAM estimates, the blue triangle for the random forest estimates, and the purple square for the actual data.

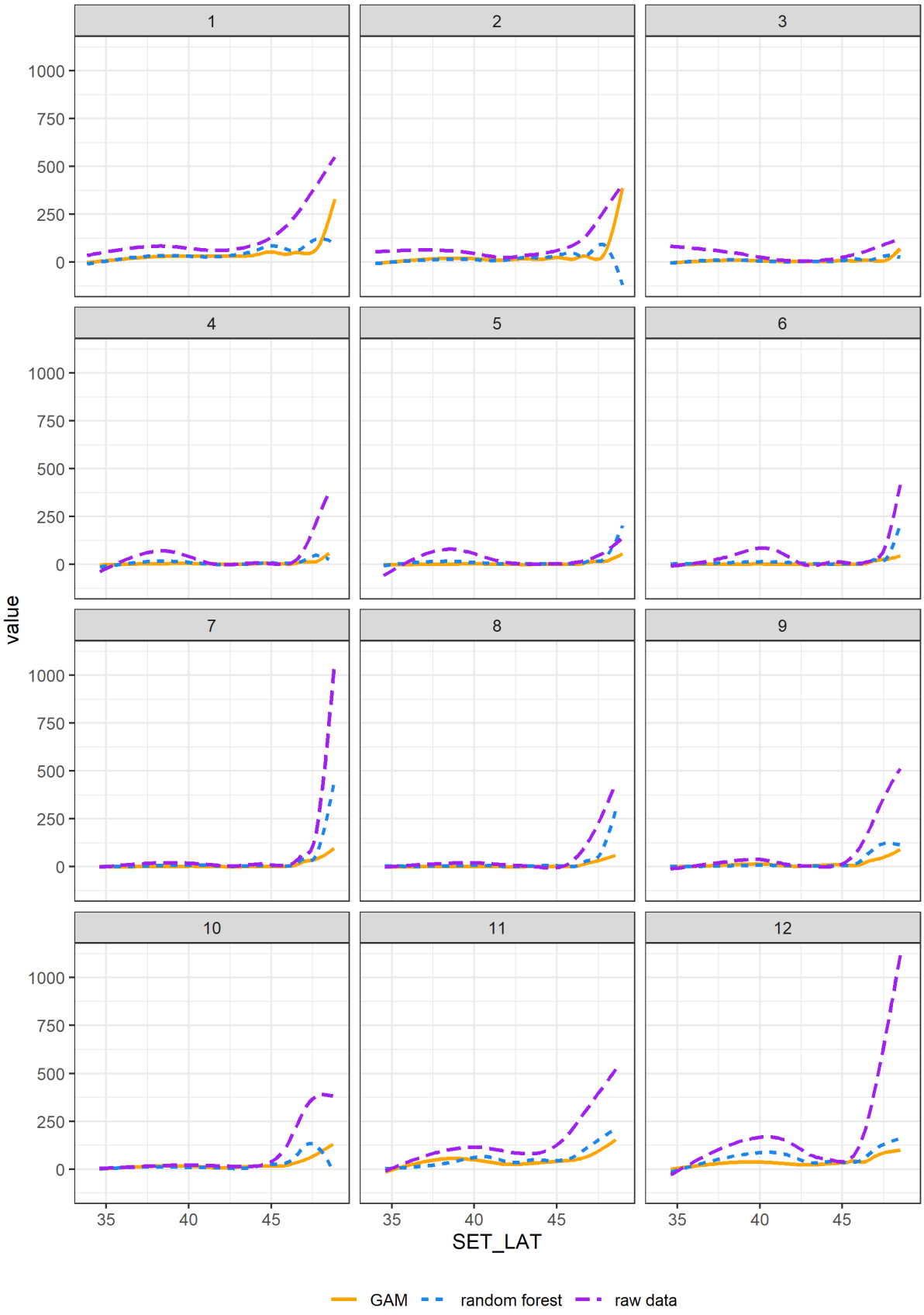



Figure 4: Smoothed mean values over latitude for the GAM and random forest model estimates of catch by haul and raw data in the WCGOP observed bottom trawl. The curves are produced with ggplot2's default geom_smooth method, which is also an mgcv GAM.

# Discussion

The STAT concluded, based on the results of their GLM and analysis of raw data concluded that it was "difficult to make a defensible argument for a value of WCGBT Survey catchability being lower than that estimated in the assessment model (0.586)."

I admit to not understanding how the SSC would use this information to recommend a prior for q. To make plain my thinking, the logic of the approach taken here is that q would be closest to 1 when the fish are at highest density/most available to the gear type. For a q greater than 1, behaviors like herding caused by the trawl net would have to occur. If the fishery-dependent data could identify the time of highest density/availability, then that would be the denominator to compare against.

The time of year when the fish are most dense/available is a different question than a simple comparison of survey season vs. non-survey season. The timeframe of highest density/availability might be something shorter than the months not surveyed.

Again, perhaps I am misunderstanding, yet a survey q of 0.586 implies that the density/availability of dogfish in the survey area would only be ~1.7 times greater if the population were ever fully available to the survey. The results shown here raise the possibility that dogfish are many times more available to bottom trawl gear during Nov-Jan than during the survey months. Taking the lower estimates from the random forest models, the simple average across Nov-Jan is 5.9 times (i.e. if survey tows occurred in those months instead of the month they actually ocurred, the random forest model estimates the average catch would be 5.9 times higher.).

Fishery-dependent data is noisy, and the model fits here show that. It is unlikely that it does tell us when the time of highest density/availabiliy of dogfish is with any certainty. Indeed, with the exception of the presence/absence regression tree model, the models show low R-squared and Deviance Explained values. However, and without putting much stock into the precision of the modeled catch amounts shown here, the relative pattern seen in Nov-Jan is robust across both the GAM and random forest models. The models do factor in some variation in space, depth, and time but of course not perfectly. And a pattern sustained over a three month period provides more information than one of shorter duration.

From my perspective, it is not clear what type of fishing behavior could explain the large differences. It isn't targeting. Effort is lower during Nov-Jan but I don't see how that could explain the pattern in full. Acknowledging the large uncertainty, I do not see reason for completely ruling out the fact that density or behavior of the fish could be a major factor behind the pattern. That is what the STAT's report seems to do. Beyond the Nov-Jan pattern, the pattern of low expected catches from April to August is likewise hard to explain based on differences in fishing behavior.

Lastly, the difference in catches in Dec-Jan seems plain to the eye in the map shown in Figure 1 of the STAT's report. Catches are more broadly dispersed in the area north of Cape Mendocino than at other times of years. I also see evidence of it in their Figure 7, yet it is harder to see.

# Appendix – Model summaries/output

With more time, I would have done a better job of describing model fits, etc. I explored variations on both model types but with limited time this is barebones output meant to provide some info on model fit, etc. With the limited time available, I do not attempt to describe GAMs and random forests. I would offer *Introduction to Statistical Learning* as a good starting points. The Second Edition is just recently published and the pdf available to download for free here: https://www.statlearning.com/.

## GAM

GAMs present a number of modeling choices, including which predictor terms should be smooth, to the type of basis k, the dimension of the basis used, and more. I attempted a number of variations (e.g. modeling year

with bs="re", which is akin to a random effect). The results in terms of Deviance Explained and predicted catches were consistent among models (i.e. Deviance Explained only varies from ~18% to ~25%).

**Binomial/classification**

```
Family: binomial
Link function: logit

Formula:
pos_dsrk ~ s(X, Y) + s(set_year) + set_month_fct + s(SET_DEPTH) +
    s(HAUL_DURATION)

Parametric coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.235588   0.009144 -25.764  < 2e-16 ***
set_month_fct.L  -0.075803   0.033554  -2.259 0.023874 *
set_month_fct.Q   3.124802   0.039092  79.935  < 2e-16 ***
set_month_fct.C  -0.079575   0.031317  -2.541 0.011055 *
set_month_fct^4  -0.211570   0.030976  -6.830 8.48e-12 ***
set_month_fct^5  -0.189649   0.029589  -6.409 1.46e-10 ***
set_month_fct^6  -0.116843   0.028736  -4.066 4.78e-05 ***
set_month_fct^7  -0.122158   0.027620  -4.423 9.74e-06 ***
set_month_fct^8  -0.099407   0.026344  -3.773 0.000161 ***
set_month_fct^9  -0.005003   0.025478  -0.196 0.844312
set_month_fct^10 -0.242434   0.024558  -9.872  < 2e-16 ***
set_month_fct^11  0.007184   0.024340   0.295 0.767890
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                   edf Ref.df Chi.sq p-value
s(X,Y)          28.623 28.989 6373.2  <2e-16 ***
s(set_year)      8.927  8.998 1538.2  <2e-16 ***
s(SET_DEPTH)     7.544  8.226 7801.2  <2e-16 ***
s(HAUL_DURATION) 7.503  8.183  871.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.288   Deviance explained = 24.1%
UBRE = 0.041712  Scale est. = 1         n = 94243
```

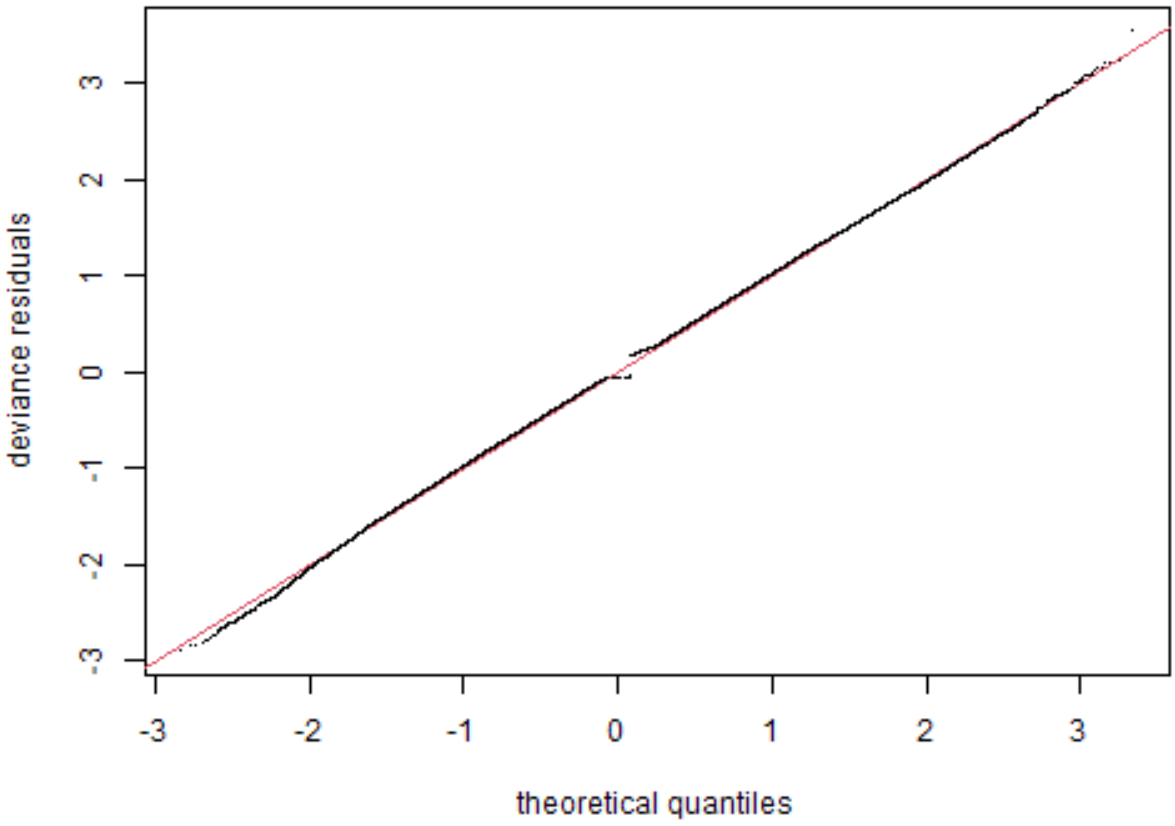
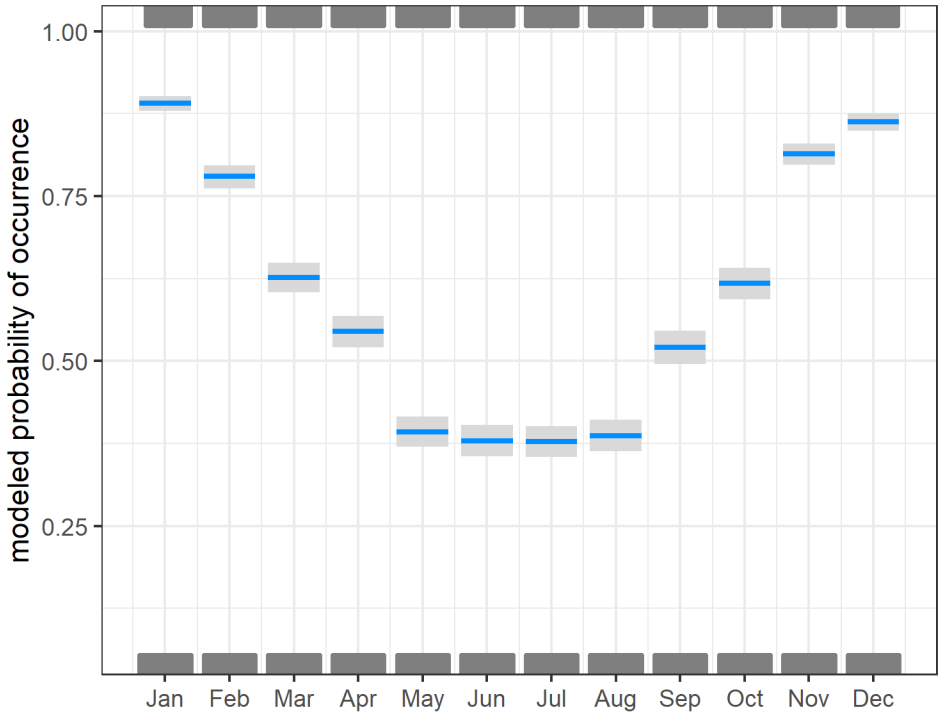



Figure 5: Residual QQ plot for binomial GAM




**Lognormal**

```
Family: gaussian
Link function: identity

Formula:
log(lbs_dsrk) ~ s(X_km, Y_km) + s(set_year) + set_month_fct +
    s(SET_DEPTH) + s(HAUL_DURATION)

Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.139194   0.008832 355.444  < 2e-16 ***
set_month_fct.L   0.515514   0.032974  15.634  < 2e-16 ***
set_month_fct.Q   2.120625   0.040316  52.600  < 2e-16 ***
set_month_fct.C  -0.600362   0.031785 -18.889  < 2e-16 ***
set_month_fct^4  -0.162067   0.032327  -5.013 5.37e-07 ***
set_month_fct^5  -0.058738   0.031064  -1.891  0.05865 .
set_month_fct^6  -0.399912   0.030937 -12.927  < 2e-16 ***
set_month_fct^7  -0.070763   0.030931  -2.288  0.02216 *
set_month_fct^8   0.081701   0.030662   2.665  0.00771 **
set_month_fct^9  -0.037612   0.030387  -1.238  0.21580
set_month_fct^10 -0.252248   0.030255  -8.337  < 2e-16 ***
set_month_fct^11 -0.063196   0.030749  -2.055  0.03986 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                   edf Ref.df      F p-value
```

```
s(X_km,Y_km)     28.518 28.982  85.76  <2e-16 ***
s(set_year)       8.727  8.976 148.06  <2e-16 ***
s(SET_DEPTH)      8.225  8.787 481.93  <2e-16 ***
s(HAUL_DURATION)  8.629  8.942  94.79  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.212   Deviance explained = 21.3%
GCV = 3.2023  Scale est. = 3.1972    n = 41182
```

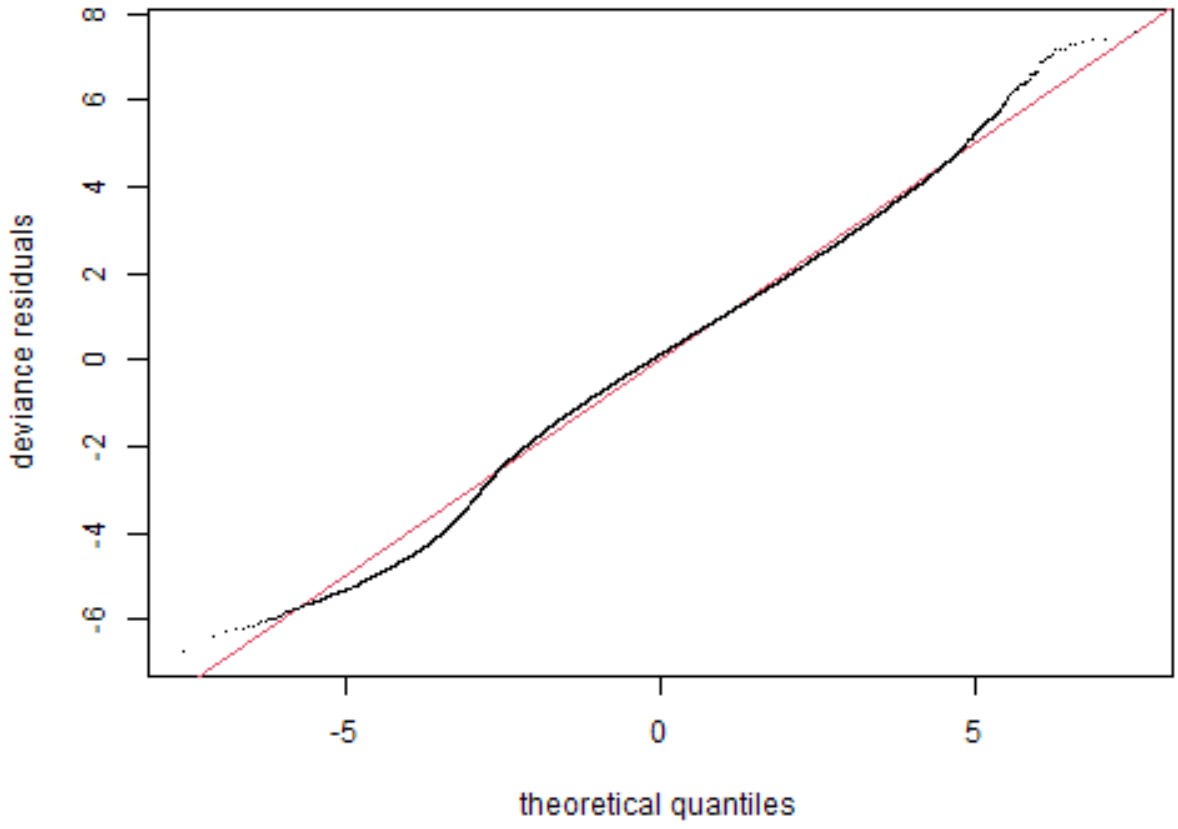



Figure 6: Figure XX: Residual QQ plot for lognormal GAM

## Random Forest

To explore the sensitivity to tuning parameters, I followed the testing, training, validation approach described in this tidymodels case study: https://www.tidymodels.org/start/case-study/.

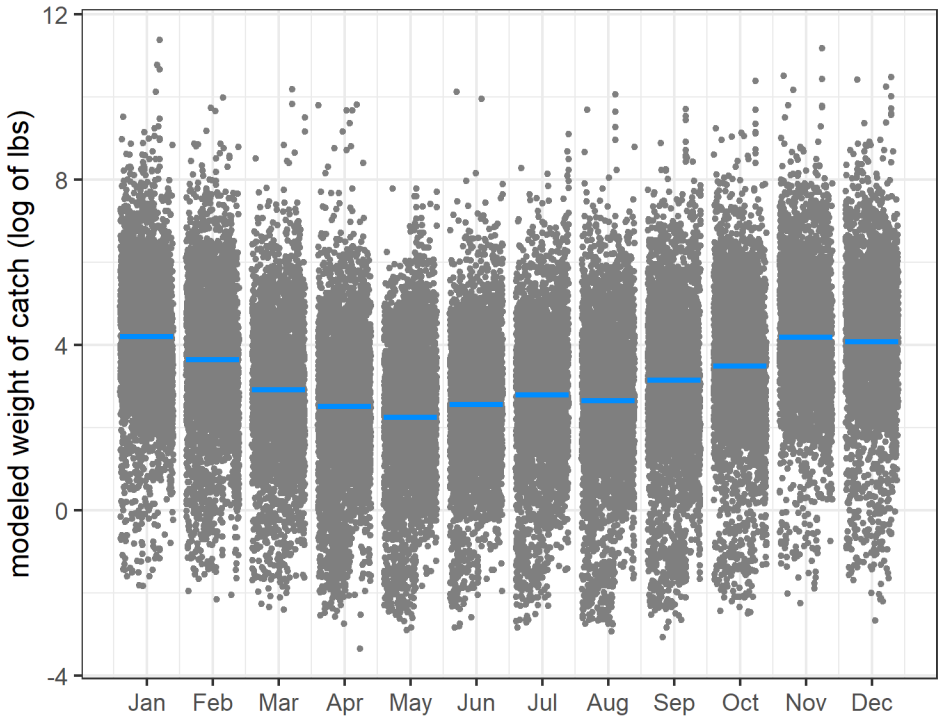



Figure 7: Conditional plot of estimated catch weight (log(lbs)) by month

### Classification

The final model AUC_ROC was 0.866. The AUC_ROC score only varied to the fourth decimal place around multiple combinations of the mtry and min.node.size parameters.

```
Call:
 ranger::ranger(x = maybe_data_frame(x), y = y, mtry = min_cols(~mtry_binom_best,      x), num.trees =

Type:                             Probability estimation
Number of trees:                  3000
Sample size:                      70681
Number of independent variables:  6
Mtry:                             2
Target node size:                 25
Variable importance mode:         impurity
Splitrule:                        gini
OOB prediction error (Brier s.):  0.148081
```

### Random Forest - Regression

This model too was insenstive to variations on the mtry and min.node.size parameters.

```
Call:
 ranger::ranger(x = maybe_data_frame(x), y = y, mtry = min_cols(~2,      x), num.trees = ~3000, min.node
```

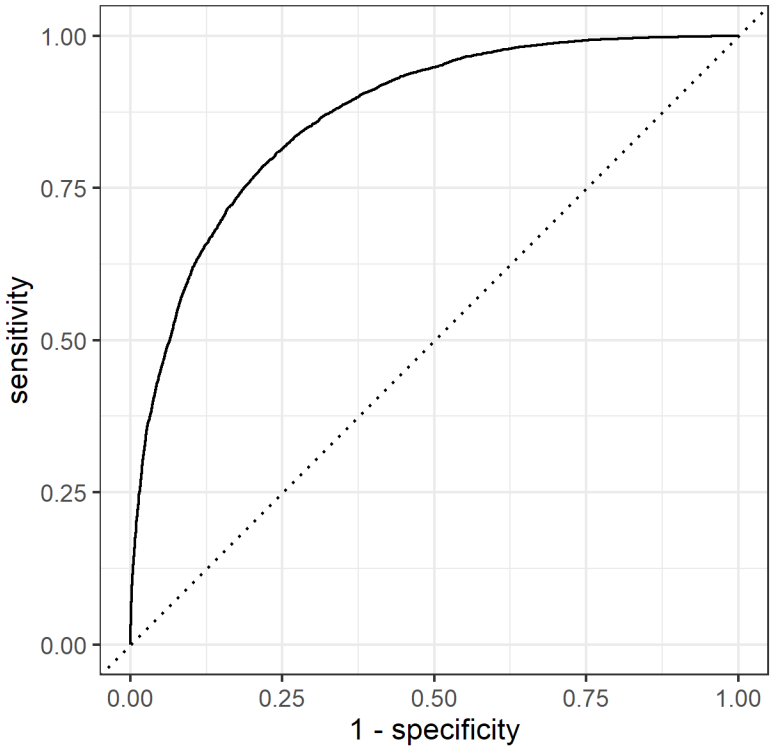



Figure 8: ROC curve for the random forest classification model

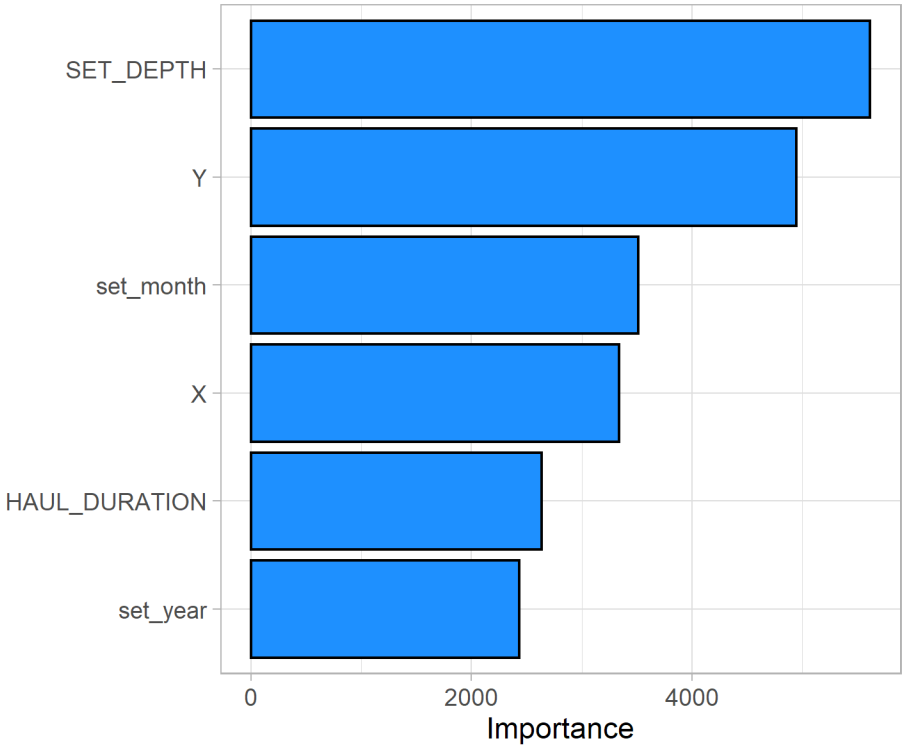



Figure 9: Variable importance plot for the random forest classification model

```
Type:                             Regression
Number of trees:                  3000
Sample size:                      30886
Number of independent variables:  6
Mtry:                             2
Target node size:                 5
Variable importance mode:         impurity
Splitrule:                        variance
OOB prediction error (MSE):       2.504774
R squared (OOB):                  0.3836585
```

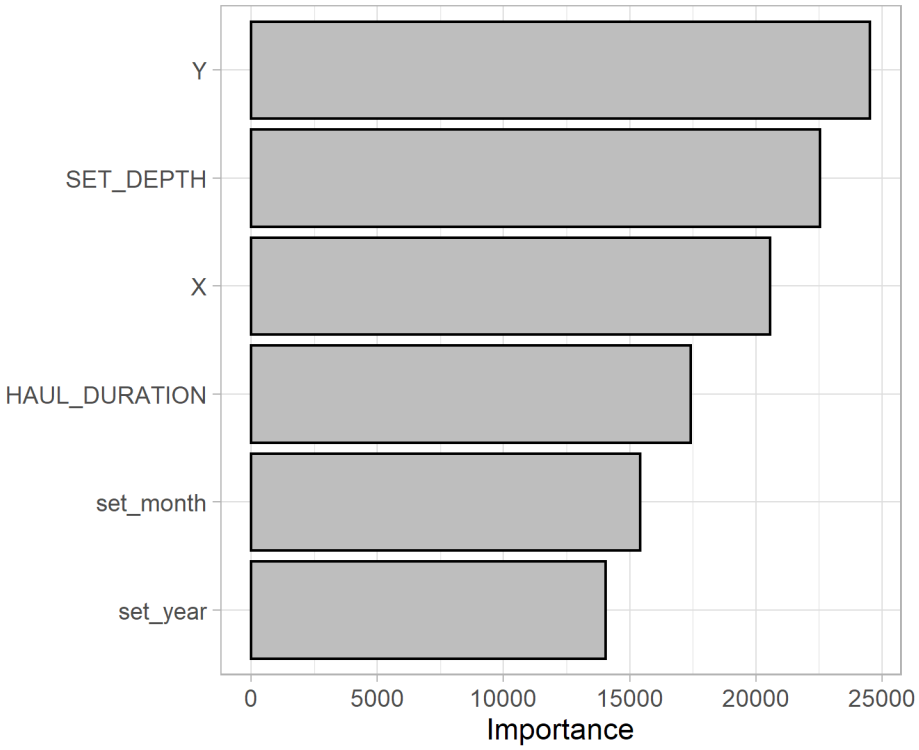



Figure 10: Variable importance plot for the random forest regression model