Estimates of Steelhead in Dungeness River

Using Sonar

Kevin See[1,*]

June 03, 2022

Contents

[1] Washington Department of Fish & Wildlife

[*] Correspondence: Kevin See <Kevin.See@dfw.wa.gov>

1 Goals

To estimate the number of adult steelhead spawners in the Dungeness river, with appropriate uncertainty, based upon a sonar system set up in the lower river. This involves several sub-tasks:

1. Expand 30 min observations into 60 min estimates. Most hours only the first 30 min of video was analyzed, but for a subset of hours the entire hour was analyzed. Develop a crosswalk between the first 30 min and the entire hour, and then predict the total counts from all the hours when only the first 30 min are available, with uncertainty from this crosswalk.
2. Exclude bull trout from images based on species composition data.
3. Account for outages. There are periods when the sonar unit was not functioning for a variety of reasons. Develop a method to interpolate across those periods of missing data.

2 Methods

2.1 Expanding 30 min to 60 min

We started by examining the counts during the first 30 minutes and the entire hour. First, we filtered for records with a confidence level of 1 (extremely confident) and a length greater than 67 cm to ensure we were only comparing records of steelhead. For each hour with both halves recorded and for each half hour, we summed the fish determined to be moving upstream, and those moving downstream, and calculated the number of net upstream fish (upstream - downstream). We then added the two half hours together to provide a number of the net upstream fish for that hour.

We compared the first half hour with the entire hour at several temporal scales. We started with a single hour, then also summed net upstream fish in 6 hour blocks, and finally summed the net upstream fish by date. The data was structured such that hours where both half hours were examined were usually consecutive at least up to the day scale, meaning each 6 hour block and each day had nearly identical amounts of time with two half hours to other periods at the same temporal scale.

For each temporal scale grouping, we fit a linear model with the counts of net upstream fish in the first 30 minutes as the independent variable and the total net upstream fish for the hour as the dependent variable. In each model, we fixed the intercept at 0, to ensure that if no fish were counted in the first half hour, we would expand that to zero fish for the entire hour. We focused on the estimated slope, hypothesizing that it should be 2.

2.2 Excluding Bull Trout

Bull trout are swimming past the sonar unit as well as steelhead, and we need to parse which fish identified by the sonar are steelhead, and exclude any bull trout. The largest bull trout sampled in any species composition data was 67 cm, so we are assuming any fish larger than 67 cm detected on the sonar is a steelhead. It then remains to filter the fish equal to or less than 67 cm long from the sonar and determine what proportion of those are steelhead, and what are bull trout.

We only have species composition data for one year, 2021. It was collected weekly, using tangle nets just upstream of the sonar location. For every fish caught, we know the date and fork length of that fish. Based on this data, we have also determined that the steelhead run on the Dungeness is over by June 1. Therefore, we have only made predictions for fish detected prior to June 1st.

We have several options to model the probability of any particular fish, less than or equal to 67 cm, being a steelhead. we could model that probability as a factor of date (perhaps with a quadratic term to capture non-linearity), or as a factor of fork length, or both. If we use date, we can interpret the probability of being a steelhead as the proportion of all fish detected on that date that are steelhead. If we use fork length (or

date and fork length), we can assign a probability to every fish detected by sonar, and assume that all fish with a probability greater than some threshold (probably 50%) are steelhead.

We chose to use fork length and the Juilan day of capture. From the species composition netting, there are 71 fish to use in this model. These can be seen in Figures 1 and 2. Since we only care about differentiating steelhead, we grouped resident rainbows with bull trout, and then fit a binomial GAM with a logit link, using splines of fork length and Julian day to predict the probability of a fish being a steelhead. We did not restrict the dataset to fish with fork lengths less than 67 cm, because larger fish have information about the shape of the logistic curve.

After fitting this GAM, we predicted the probability of being a steelhead for all fish observed on the sonar that were smaller than or equal to 67 cm, based on their length and Julian day of observation. Any fish with a probability of 50% or greater we assigned to be a steelhead. We then applied the same model (Section 2.1) to expand 30 minute counts for small fish to full hour counts. We added counts or estimates of large fish to small fish for each time period to estimate total net steelhead moving upstream for each time period. Estimates of large and small fish within the same time period were assumed to be independent when calculating the standard error.

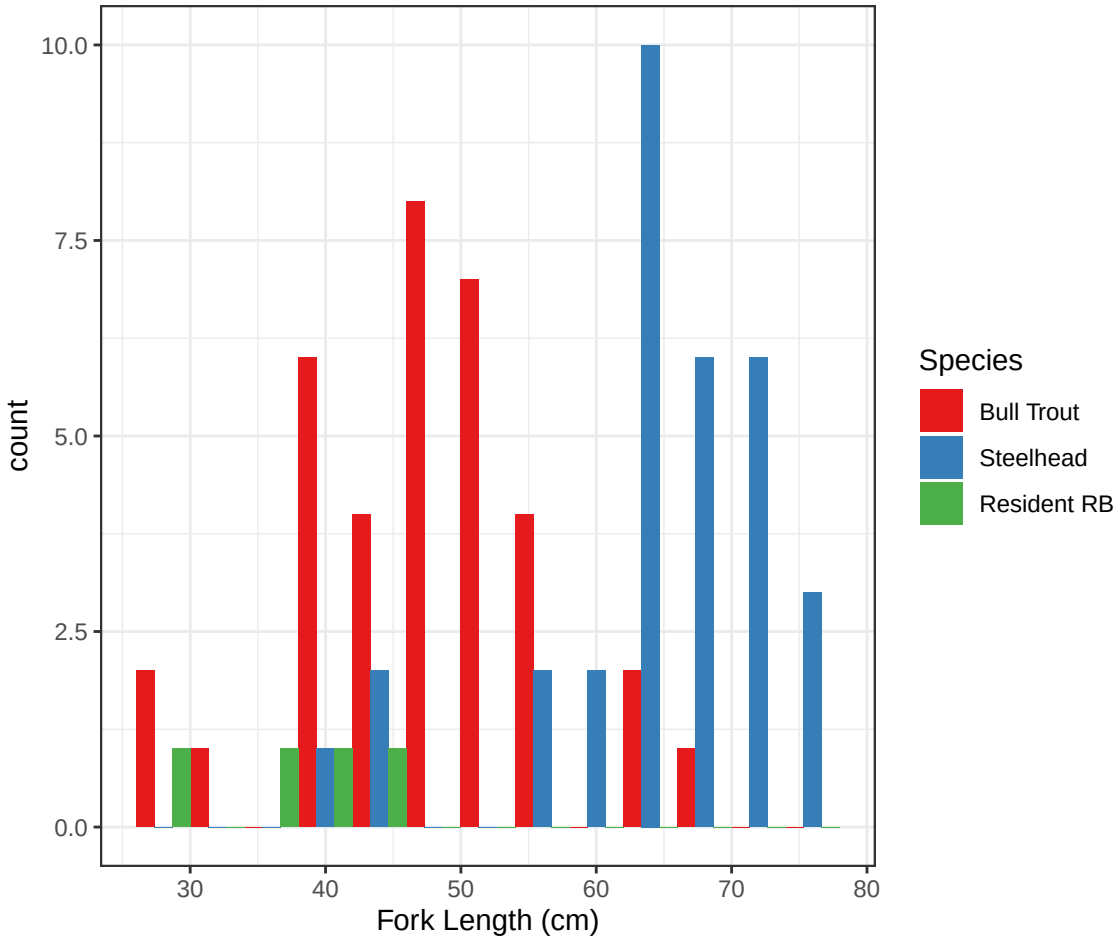

Figure 1: Histogram of forklengths, colored by species.

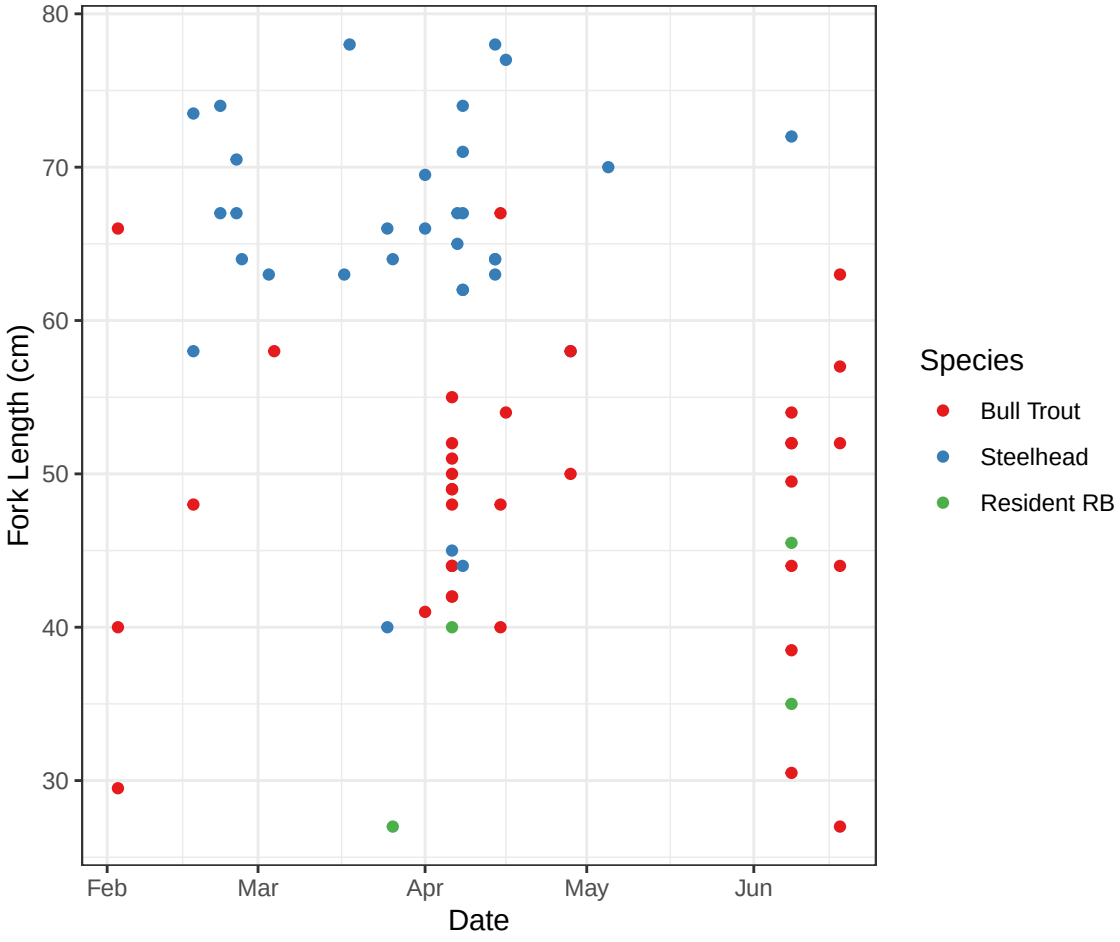

Figure 2: Scatterplot of date of capture and forklength, colored by species.

Table 1: Table showing how many periods are in each group of data, and how many of those periods are NAs (missing values).

Time Scale	Year	n Periods	n NAs	% NA
Hour	2019	2112	128	6.1
6 Hour Block	2019	352	20	5.7
Day	2019	88	5	5.7
Hour	2020	2597	509	19.6
6 Hour Block	2020	433	83	19.2
Day	2020	109	20	18.3
Hour	2021	2867	126	4.4
6 Hour Block	2021	478	10	2.1
Day	2021	120	0	0.0

2.3 Missing Data

There are periods when the sonar was not functioning, for a variety of reasons. Rather than ignore those time periods, and assume that no steelhead were passing then, we would prefer to impute net upstream fish for those missing values.

The first step is to expand the estimates of net upstream fish for periods when the sonar only partially operated (e.g. 14 hours out of a 24 hour day). We did this by dividing the estimate for that period by the percent of time the sonar was operational in that period. This assumes that fish are behaving similarly for that entire period.

The next step is to deal with those periods when the sonar was not operating at all, where we have truly missing data. Table 1 shows how much data was missing for each year, depending on how the periods were constructed (e.g. hourly, hourly blocks, daily).

To interpolate across those periods of missing data, we employed time-series models. Using the `forecast` package in R, we fit an ARIMA (auto-regressive integrated moving average) model, and let the `auto.arima` function determine the model with the best order (number of auto-regressive, moving average and difference steps) for each year and time-scale combination. We used the uncertainty from this model (σ^2) for all predictions.

We examined several forms of interpolation across the missing data, including a Kalman filter, linear regression and moving average. The Kalman filter uses the ARIMA structure to estimate the missing data. A linear regression essentially draws a straight line from the data point prior to the first missing data and the data point after the last missing data point for each gap in the time series. A moving average approach uses two non-missing values prior to the missing data point, and two non-missing values after, weights them exponentially by their distance from the missing data point, and calculates the weighted mean.

3 Results

3.1 Expanding 30 min to 60 min

The expansion factor (i.e. slope) changes depending on the temporal scale that data is summarized on. Figure 3 shows the various regressions, comparing them with the 1-1 line and the expected slope of two. None of the temporal scales produced a slope of two, but the longer the temporal scale the closer it got to that expected value.

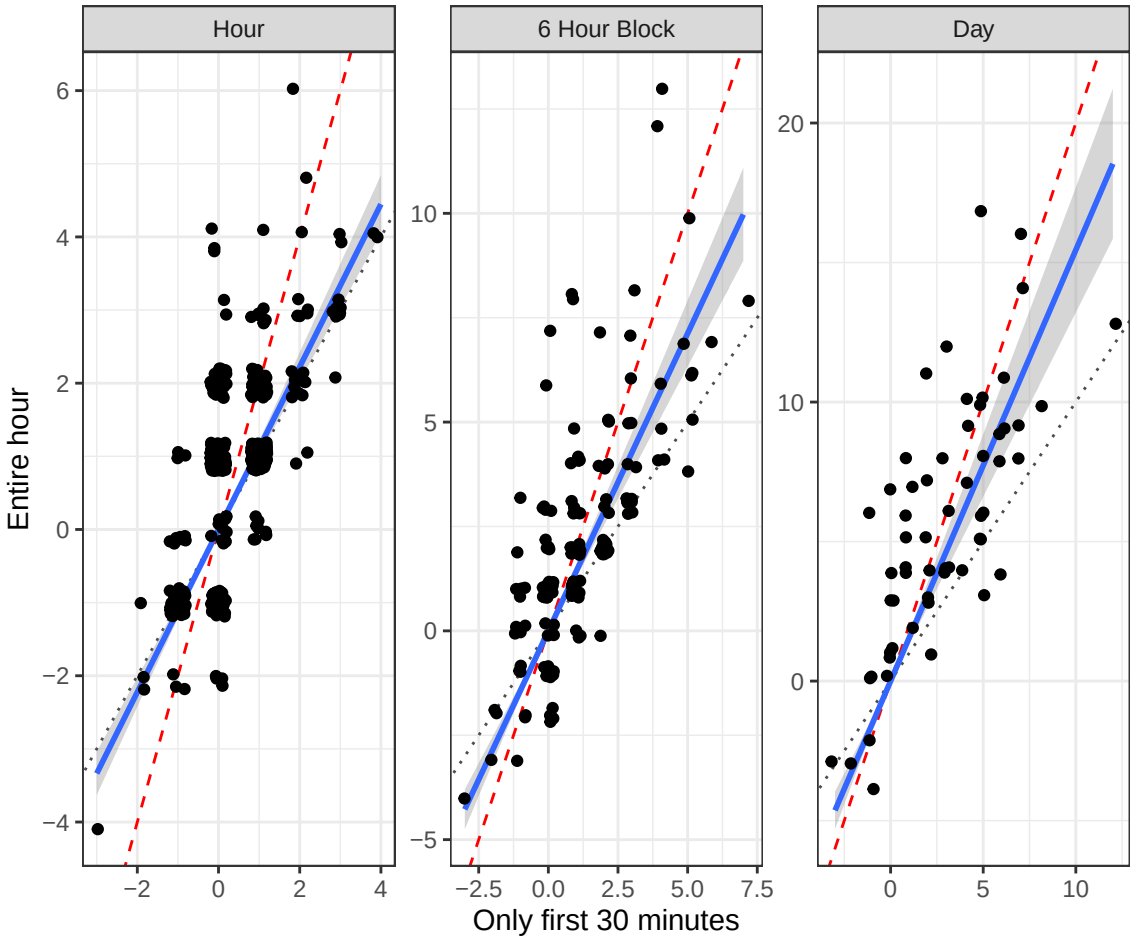

Figure 3: Scatter plots showing the net counts of fish moving upstream, using hours with both the first 30 minutes and second 30 minutes. The counts are summarized by hour, six hour blocks and entire day (24 hours) in the different facets. The dashed red line is has a slope of 2 (expected value), the dotted grey line has a slope of 1, and the blue line is the linear regression fit to that data, with 95% confidence intervals.

Table 2 shows the summary of linear models fit to data summarized at various time scales. We summarized the estimated number of steelhead larger than 67 cm at the day scale (summing estimates at smaller temporal scales) and plotted the time-series in Figure 4 to show the differences caused by summarizing data at different time scales.

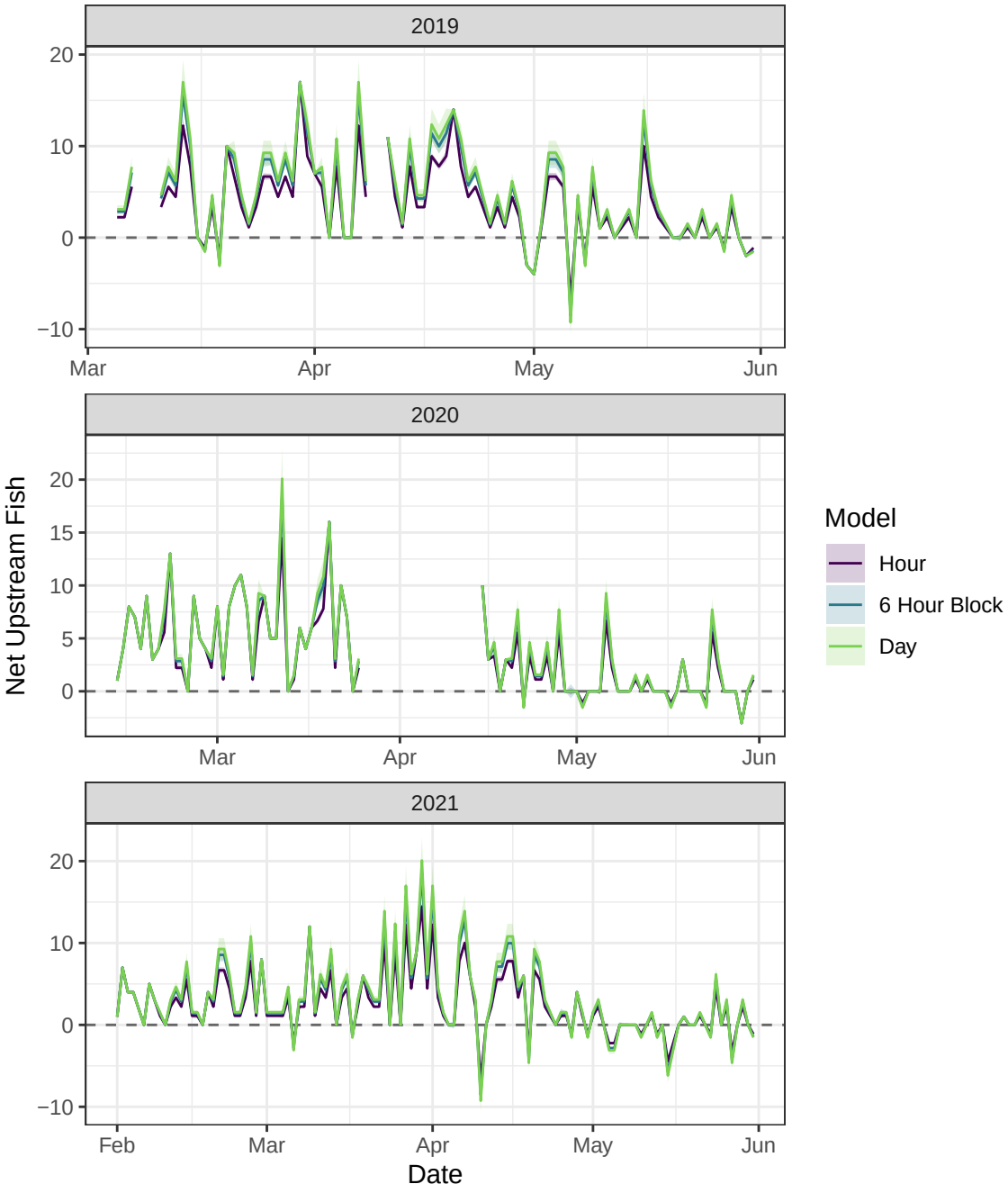

Figure 4: Time-series of estimates based on available data, using only extremely confident observations of fish greater than 67 cm, faceted by year. Colors correspond to which regression model was used to expand the 30 minutes observations. Any uncertainty shown is derived from the linear regression model.

Table 2: Results of fitting linear models with total net upstream fish as the response and the net upstream fish in the first 30 minutes as the covariate with no intercept.

Time Scale	Slope	SE	95% CI	R2
Hour	1.11	0.05	1.01-1.21	0.56
6 Hour Block	1.42	0.08	1.27-1.58	0.67
Day	1.54	0.11	1.32-1.77	0.76

Table 3: Estimates of total net upstream fish larger than 67 cm, using only extremely confident observations. There are three estimates per year, corresponding to the three different regression models for expanding 30 minute observations.

Time Scale	Year	Estimate	SE
Hour	2019	305	1.27
6 Hour Block	2019	376	2.47
Day	2019	403	4.23
Hour	2020	297	0.78
6 Hour Block	2020	324	1.48
Day	2020	334	2.50
Hour	2021	301	1.26
6 Hour Block	2021	363	2.42
Day	2021	386	4.21

3.2 Excluding Bull Trout

Figure 5 shows the fitted GAM that predicts the probability of being a steelhead based on a fish's length and date of capture. Note that the mean probability of being a steelhead for a 67 cm long fish, averaged across the entire season, would be 85.8% of being a steelhead with this model. Also note that a fish 60 cm long would not be considered a steelhead if observed at the very beginning of the season or after the beginning of May, but would if observed in late March or April.

Applying this model and rule-set to all the observed fish smaller than or equal to 67 cm, including the predictive model to expand 30 minute counts to full hour counts, a number of additional steelhead are added to our estimate each year (Table 4). The total estimates (including all fish larger than 67 cm, as well as fish less than or equal to 67 cm that are predicted to be steelhead) are shown in Table 5.

Table 4: Estimates (SE) of total net upstream steelhead smaller than 67 cm, using only extremely confident observations. There are three estimates per year, corresponding to the three different regression models for expanding 30 minute observations.

Time Scale	2019	2020	2021
Hour	73 (0.5)	150 (0.6)	42 (0.7)
6 Hour Block	88 (0.9)	167 (1.1)	51 (1.2)
Day	94 (1.6)	173 (1.9)	54 (1.9)

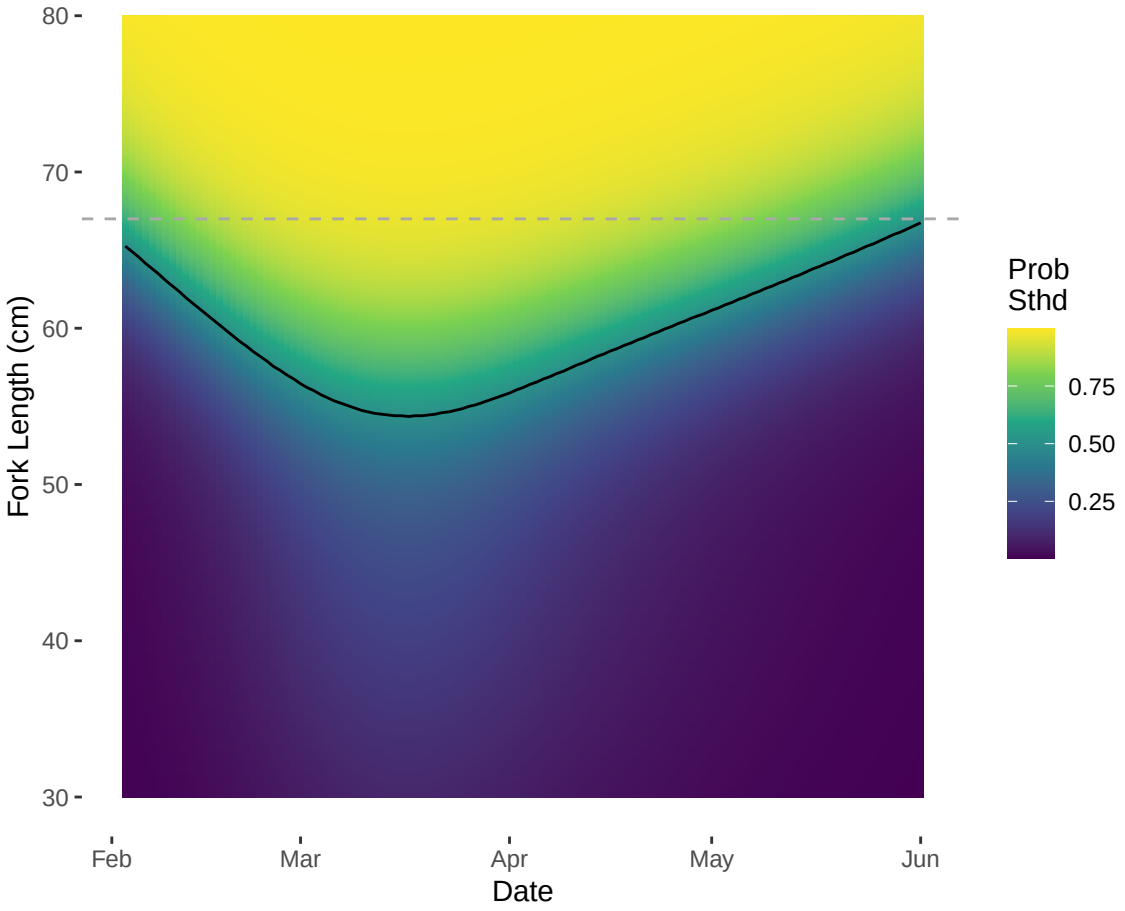

Figure 5: The color depicts the probability of fish being a steelhead given the date of capture and fork length. Fish above the black line would have a greater than 50% probability of being a steelhead. The dotted line shows the 67 cm threshold for which fish we will be applying this model to.

Table 5: Estimates (SE) of total net upstream steelhead, using only extremely confident observations. There are three estimates per year, corresponding to the three different regression models for expanding 30 minute observations.

Time Scale	2019	2020	2021
Hour	377 (1.4)	447 (1)	343 (1.4)
6 Hour Block	464 (2.6)	491 (1.9)	414 (2.7)
Day	497 (4.5)	508 (3.1)	441 (4.6)

3.3 Missing Data

Figure 6 shows the periods when the sonar array was not operating, and Figure 7 shows how that impacts the time-series of fish counts. Note the large period in 2020 when the sonar was shut down due to COVID-19.

As the temporal scale on which counts are aggregated increases, the amount of missing data decreases. For example, if three hours are missing within a day, we can expand the rest of the day's counts by the percent of time the sonar was operational, so that day will not be "missing" at the day time-scale, although those three hours still are if we are operating on an hour time-scale.

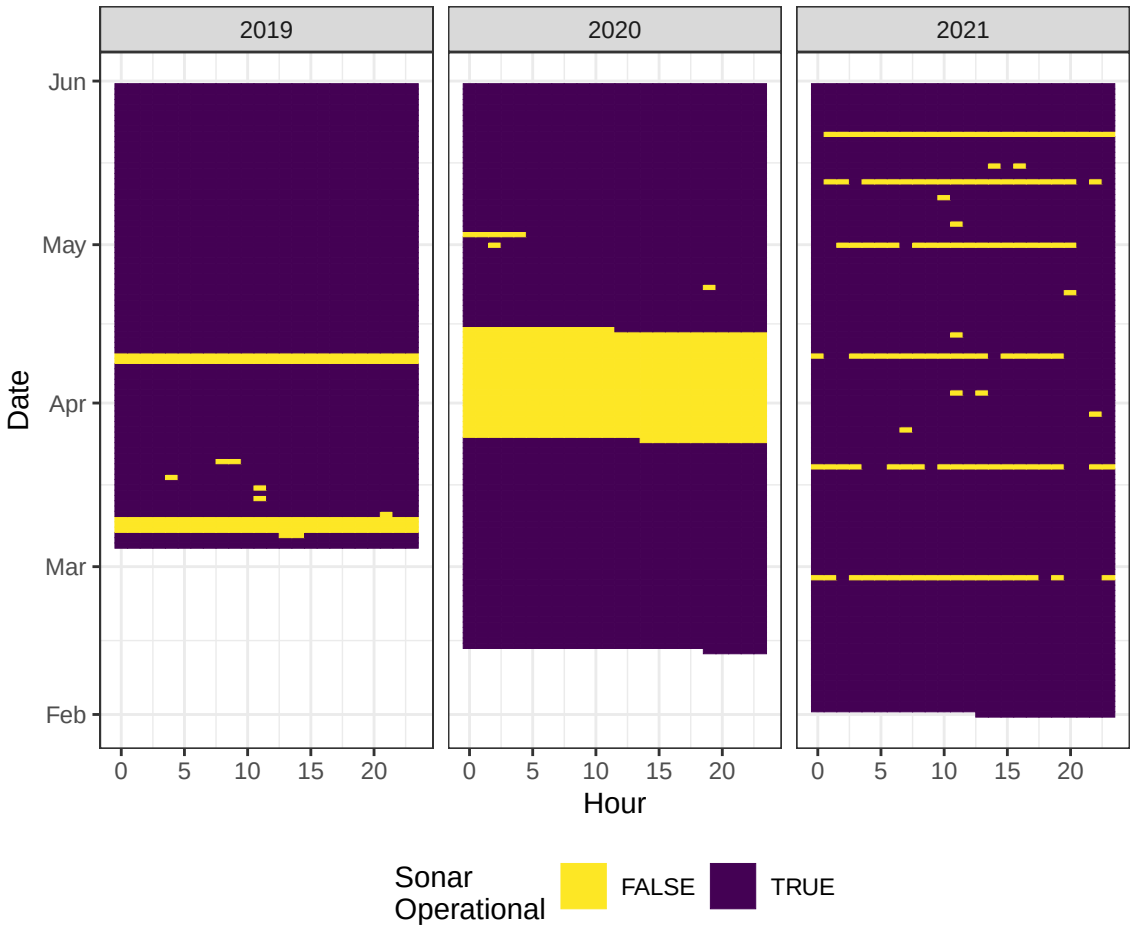

Figure 6: Purple depicts hours when the sonar was working, while yellow indicates the sonar was not functioning.

After interpolating across the missing data, Table 6 displays how many fish were added to each year's estimate, based on the temporal scale and the interpolation method. Table 7 provides final estimates of total net upstream steelhead each year, including fish smaller than 67 cm and periods of missing data, split out by the temporal scale the data was summarized on.

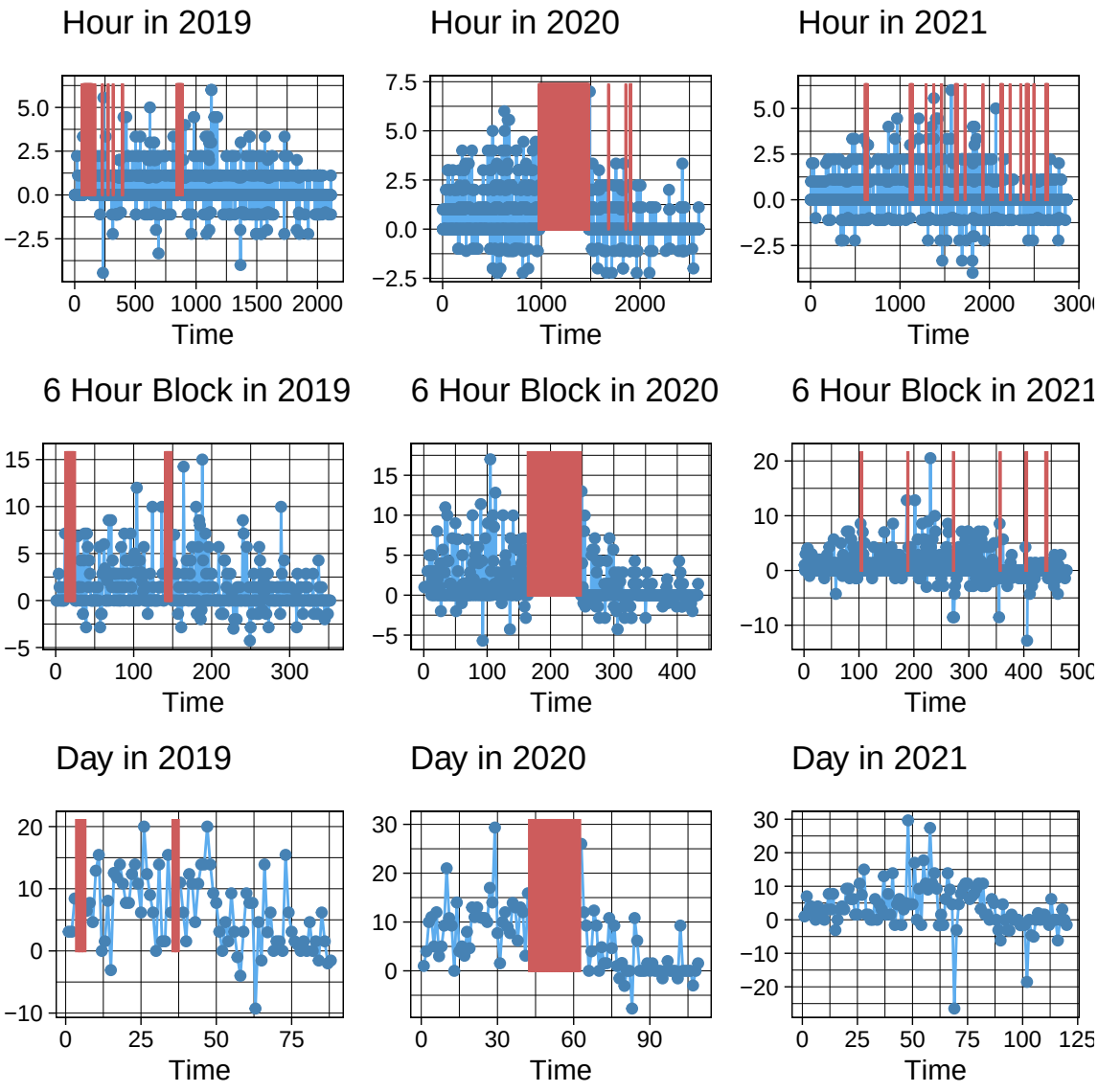

Figure 7: Time series of upstream fish in blue, with missing data highlighted in red.

Table 6: Estimates (SE) of how many net upstream steelhead are added to the totals from periods with wholly missing data. Interpolation methods include the Kalman filter, linear model, and moving average. There are three estimates per year, corresponding to the three different temporal scales. Blank cells indicate no interpolation was necessary for that year / temporal scale combination.

Time Scale	Year	Kalman	Linear	MA
Hour	2019	26 (8.6)	93 (8.6)	83 (8.6)
6 Hour Block	2019	38 (11.8)	70 (11.8)	46 (11.8)
Day	2019	32 (12.2)	39 (12.2)	36 (12.2)
Hour	2020	42 (18.7)	1 (18.7)	-1 (18.7)
6 Hour Block	2020	236 (23.9)	0 (23.9)	217 (23.9)
Day	2020	205 (24.1)	419 (24.1)	327 (24.1)
Hour	2021	3 (8)	3 (8)	-1 (8)
6 Hour Block	2021	4 (8.3)	-2 (8.3)	2 (8.3)
Day	2021	-	-	-

Table 7: Estimates (SE) of total net upstream steelhead, using only extremely confident observations, and after interpolating counts for periods of missing data. Interpolation methods include no interpolation, the Kalman filter, linear model, and moving average. There are three estimates per year, corresponding to the three different temporal scales.

Time Scale	Year	None	Kalman	Linear	MA
Hour	2019	377 (1.4)	403 (8.7)	471 (8.7)	460 (8.7)
6 Hour Block	2019	466 (2.6)	504 (12.1)	536 (12.1)	512 (12.1)
Day	2019	500 (4.5)	533 (13)	540 (13)	536 (13)
Hour	2020	447 (1)	489 (18.8)	447 (18.8)	446 (18.8)
6 Hour Block	2020	491 (1.9)	727 (24)	491 (24)	709 (24)
Day	2020	528 (3.1)	733 (24.3)	947 (24.3)	855 (24.3)
Hour	2021	343 (1.4)	346 (8.1)	347 (8.1)	342 (8.1)
6 Hour Block	2021	418 (2.7)	422 (8.8)	415 (8.8)	420 (8.8)
Day	2021	433 (4.6)	-	-	-

4 Discussion Points

- What to do with rows where `data_recorded` is "Partial?" This includes 175 rows, or 1.8% of the data. Exclude and treat as missing data? Or is there a better way to parse this? Currently I've filtered it out and treated it as missing.
- What should we do with observations with confidence of 2 or 3? They are currently excluded completely.
- What to do about kelts?

4.1 Expanding 30 min to 60 min

- The regression between counts in the first hour and the entire hour shows a consistent expectation that the counts in the second part of the hour will be less than counts in the first part. This holds regardless of whether we aggregate data by hour, day or something in between.
- Currently I'm calculating the net upstream totals for each half hour period before running the regressions. If there's a compelling reason to either run separate regressions for downstream and upstream moving fish, or treat downstream and upstream fish from the same hour as two distinct data points, let's talk about that.
  - We could run separate regressions for each year, but if the methodology is the same year-to-year, I don't see why we'd expect different results.
  - Three years is not enough to use year as a random effect, but perhaps in a few more years we could explore this as an option.

4.2 Excluding Bull Trout

- Should we assume any fish smaller than 40 cm is a bull trout, since that was the smallest steelhead length recorded? Under the current method, this doesn't matter because any fish smaller than 54.4 cm is excluded from the steelhead counts regardless of date of capture.

- I updated this model to include both the Julian day of capture and the fish length, using a spline for both covariates. The fish length spline turned out to be pretty close to a straight line, with larger fish being more likely to be a steelhead. The Julian day of capture had a peak probability of being a steelhead occurring in mid- to late-March, and tapering off on either side. Given there were a number of bull trout caught in the very beginning and towards the end of the sampling period, this shape makes sense. However, it should be noted that including Julian day of capture reduced the estimated number of small steelhead (less than 67 cm) in all years. It was a small reduction in 2020 and 2021, but a substantial one in 2019. This appears to be because there were a large number of small fish detected late in the run in 2019. Because they were so late, virtually none of those fish were predicted to be steelhead using this updated model.

4.3 Missing Data

- Any of the interpolation models we tested (Kalman filter, linear regression or moving average) resulted in larger estimates of steelhead moving upstream (Table 7), but differed in which one provide the biggest increase depending on the time-scale and year.
- The uncertainty (e.g. standard error) grew when incorporating those missing data, which is appropriate. The uncertainty grew substantially in 2020, when there was a large period of missing data due to COVID restrictions.
- Depending on how big the missing data gaps are, and the time-scale we are aggregating data on, there were some years and time-scales with no missing data (e.g. 24 hour scale in 2021). The lack of missing data relies on using the percentage of hours when sonar was operational within each time-step to increase the estimates for any time-steps when the operational time was less than 100%.

- The alternative to expanding time-steps when the sonar was partially operational is to remove all data from those time-steps and treat them as missing data.
- It's unclear to me whether that would have a substantial impact on the overall estimates or uncertainty.