Estimates of Steelhead in Dungeness River

Using Sonar

Kevin See[1,*]

May 13, 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.

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.

Currently, we are only using fork length, because although the date of capture is probably available, it is not in the current data set. From the species composition netting, there are 71 fish caught with fork lengths. These can be seen in Figure 1. Since we only care about differentiating steelhead, we grouped resident rainbows with bull trout, and then fit a binomial GLM with a logit link, using the fork length 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 GLM, 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. 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 netsteelhead 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.

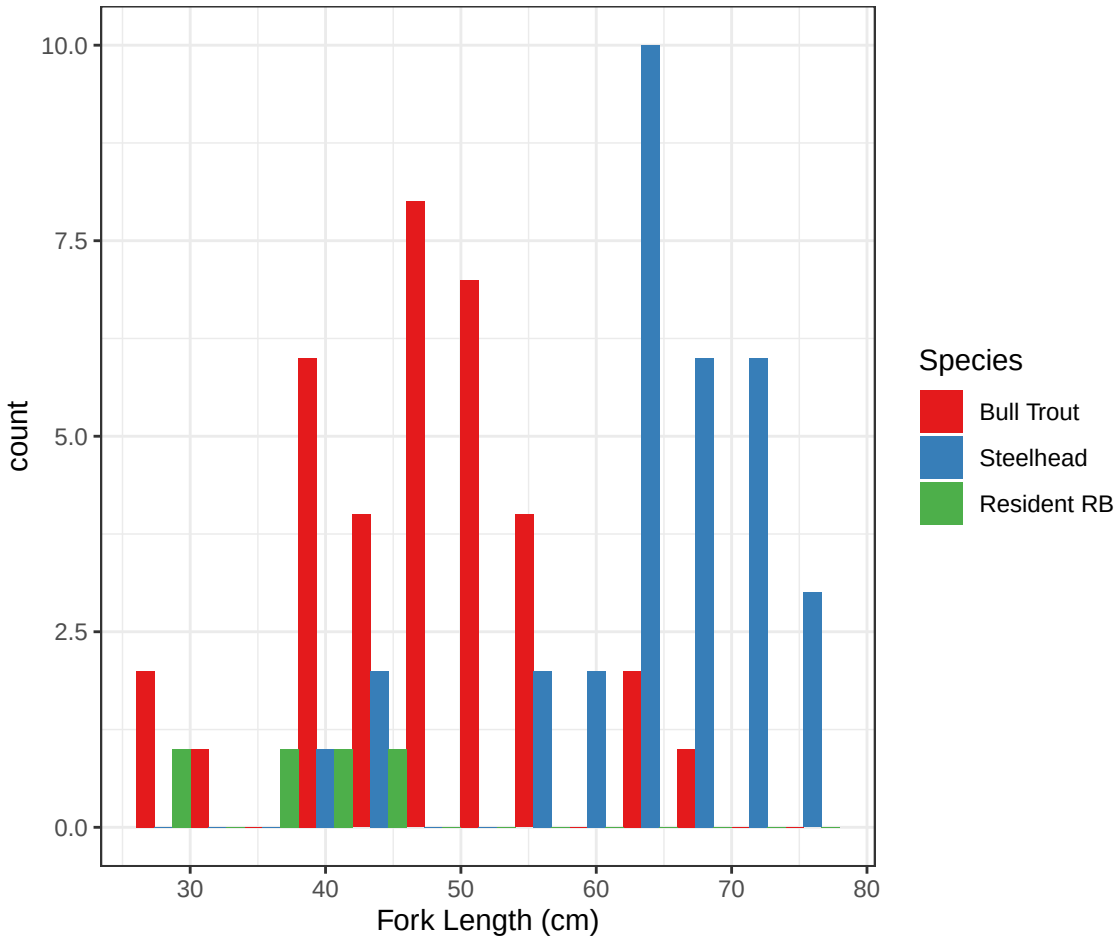

Figure 1: Histogram of forklengths, colored by species.

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

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	3240	153	4.7
6 Hour Block	2019	540	24	4.4
Day	2019	135	6	4.4
Hour	2020	3929	517	13.2
6 Hour Block	2020	655	84	12.8
Day	2020	165	20	12.1
Hour	2021	3395	131	3.9
6 Hour Block	2021	566	10	1.8
Day	2021	142	0	0.0

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 2 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.

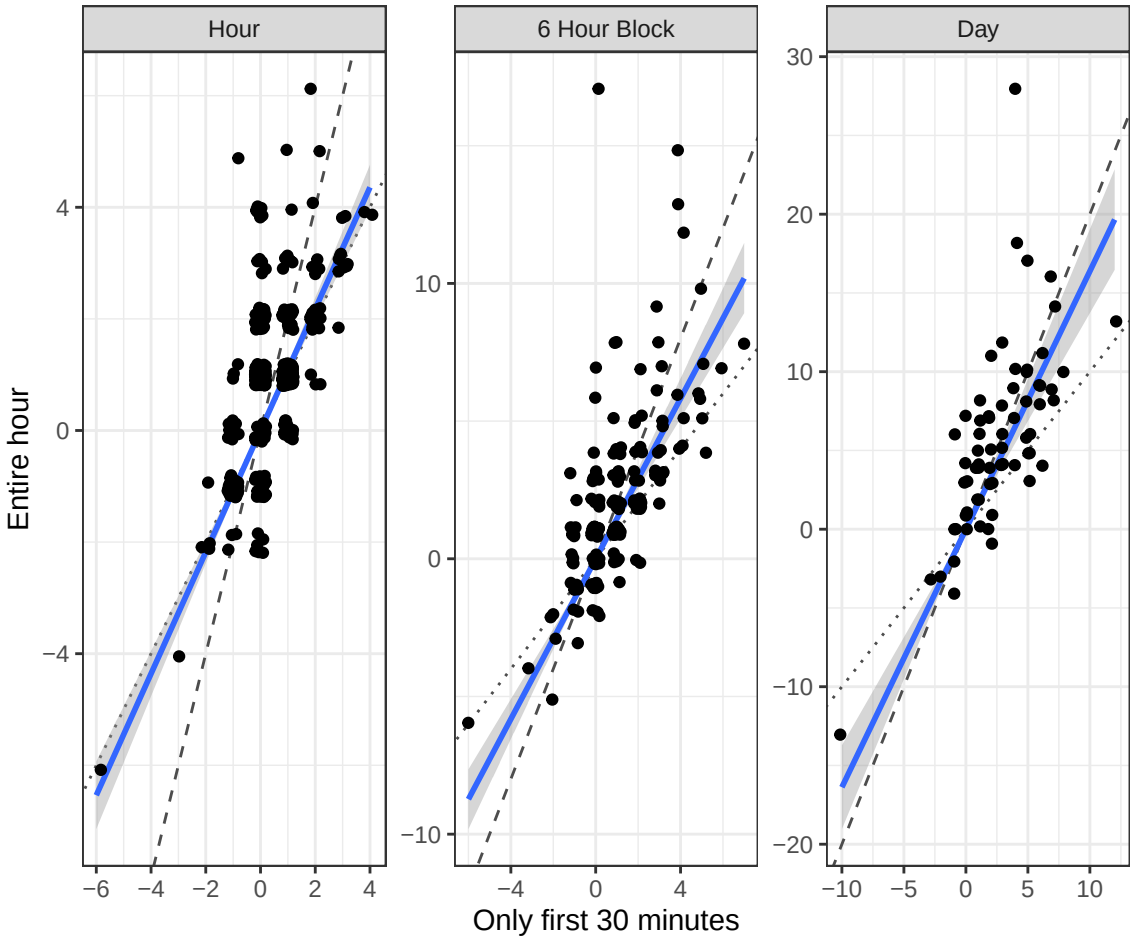

Figure 2: Scatter plots showing the net counts of fish moving upstream, using hours with both the first 30 minutes and second 30 minutes. The counts are summarized by hour, six hour blocks and entire day (24 hours) in the different facets. The dashed line is has a slope of 2 (expected value), the dotted 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 3 to show the differences caused by summarizing data at different time scales.

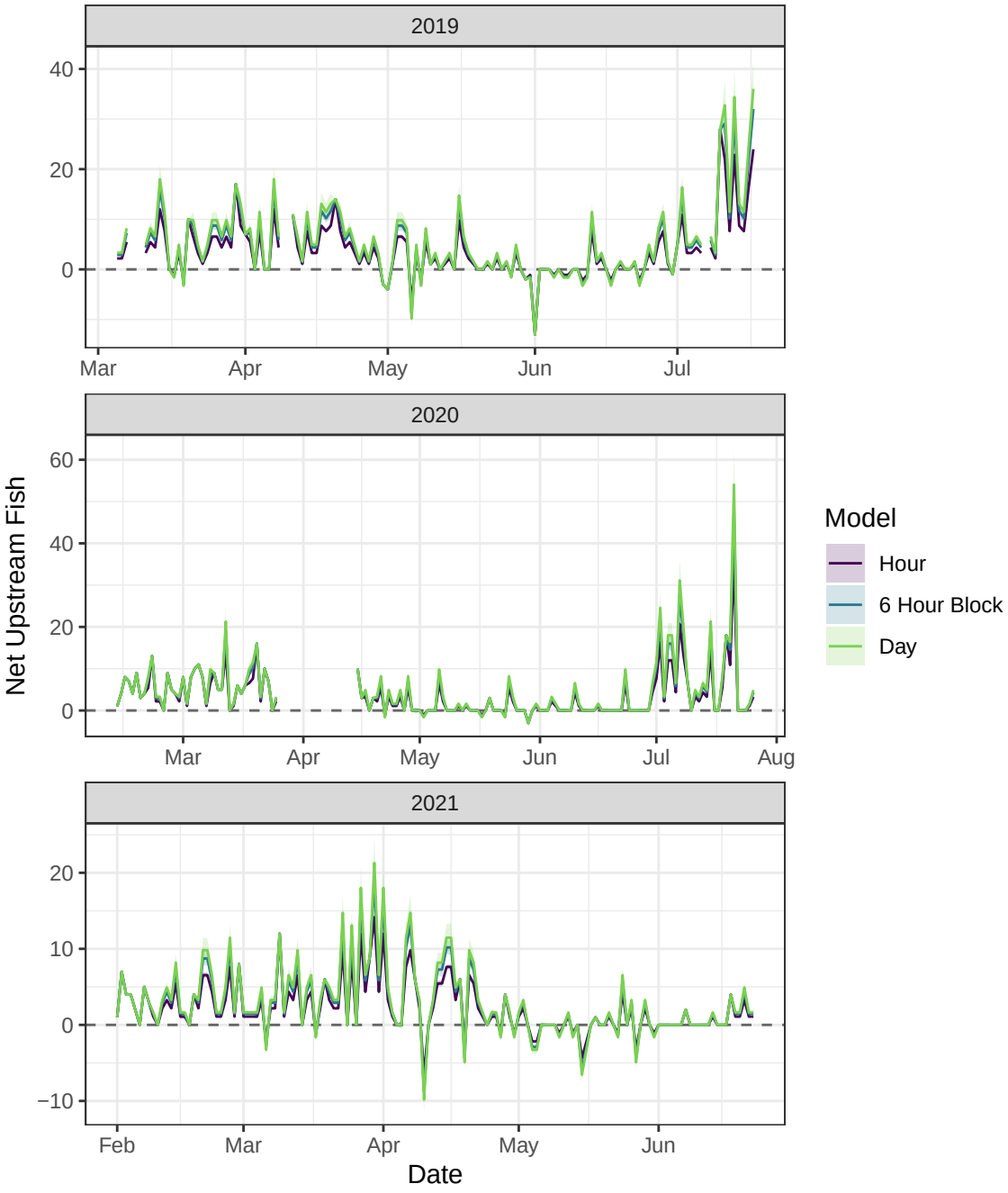

Figure 3: Time-series of estimates based on available data, using only extremely confident observations of fish greater than 67 cm, faceted by year. Colors correspond to which regression model was used to expand the 30 minutes observations. Any uncertainty shown is derived from the linear regression 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	1.09	0.05	0.99-1.19	0.51
6 Hour Block	1.46	0.09	1.27-1.64	0.58
Day	1.64	0.13	1.37-1.9	0.68

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	480	1.69
6 Hour Block	2019	617	4.08
Day	2019	686	7.85
Hour	2020	513	1.49
6 Hour Block	2020	609	3.74
Day	2020	657	7.36
Hour	2021	311	1.30
6 Hour Block	2021	386	2.79
Day	2021	424	5.04

3.2 Excluding Bull Trout

Figure 4 shows the fitted GLM that predicts the probability of being a steelhead based on a fish's length. Note that a 67 cm long fish would have a 85.6% of being a steelhead with this model.

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.

3.3 Missing Data

Figure 5 shows the periods when the sonar array was not operating, and Figure 6 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

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	258 (1.3)	169 (0.7)	51 (0.7)
6 Hour Block	333 (3.2)	192 (1.5)	61 (1.5)
Day	370 (6.7)	204 (2.5)	67 (2.5)

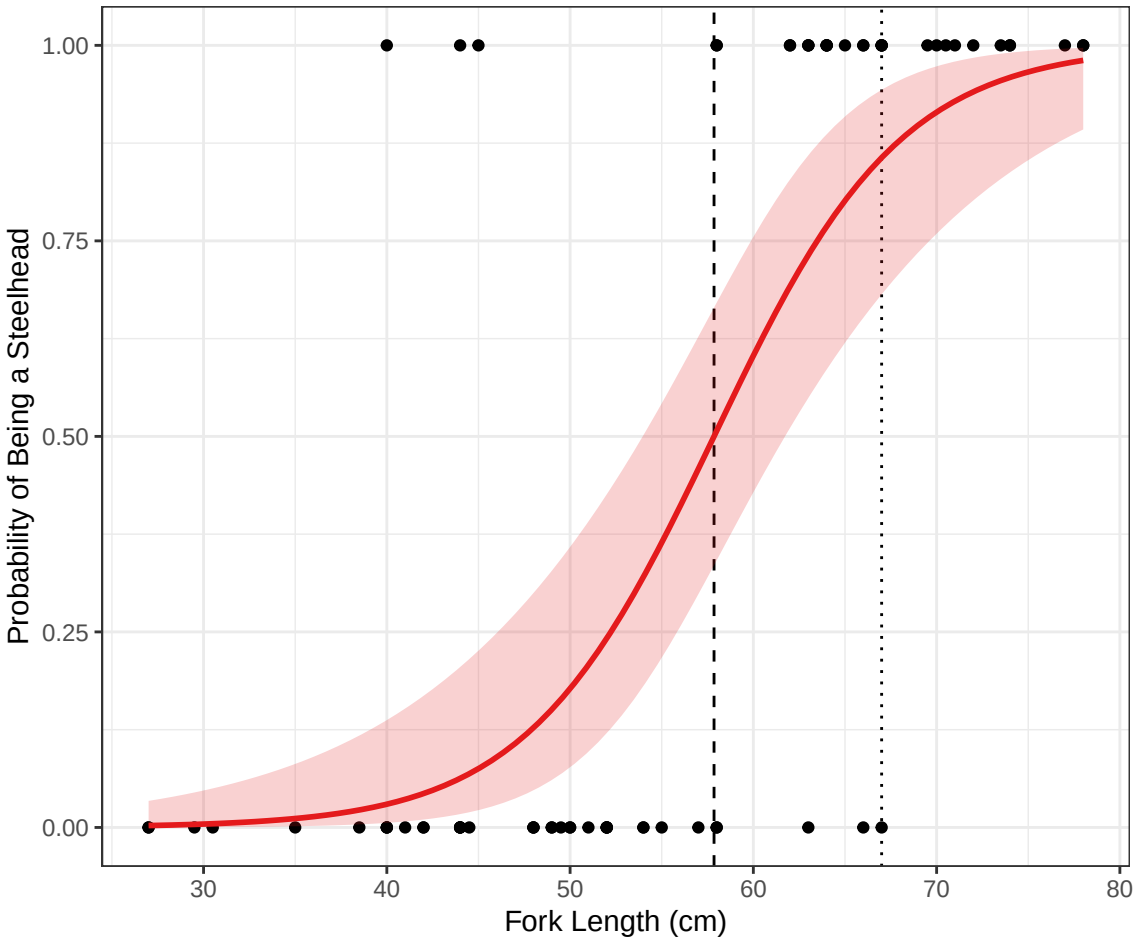

Figure 4: Points show the fork length of steelhead (along the top) and non-steelhead (along the bottom), with the fitted binomial GLM in red. The dashed line shows where fish greater than that would have a greater than 50% probability of being a steelhead. The dotted line shows the 67 cm threshold for 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	738 (2.2)	681 (1.7)	362 (1.5)
6 Hour Block	950 (5.2)	801 (4)	448 (3.2)
Day	1056 (10.3)	861 (7.8)	490 (5.6)

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.

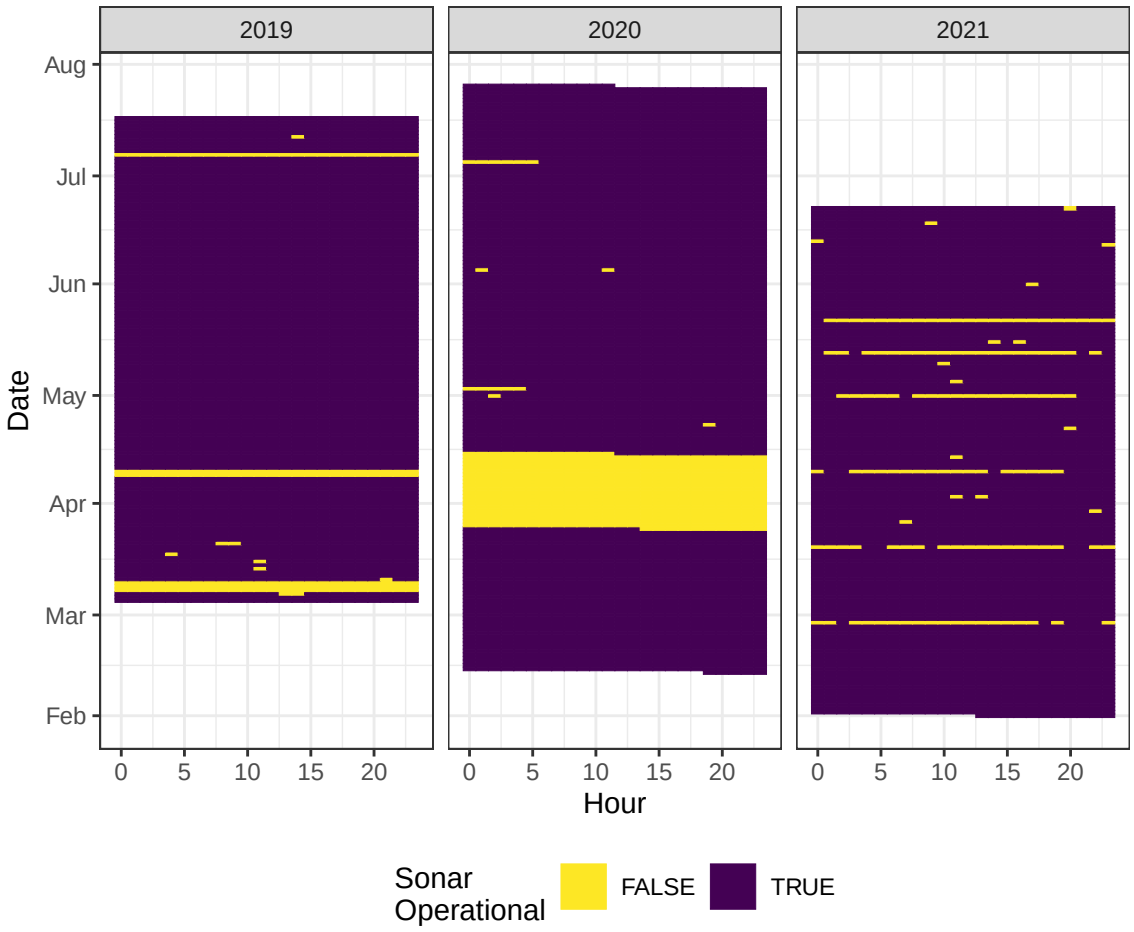

Figure 5: Purple depicts hours when the sonar was working, while yellow indicates the sonar was not functioning.

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.

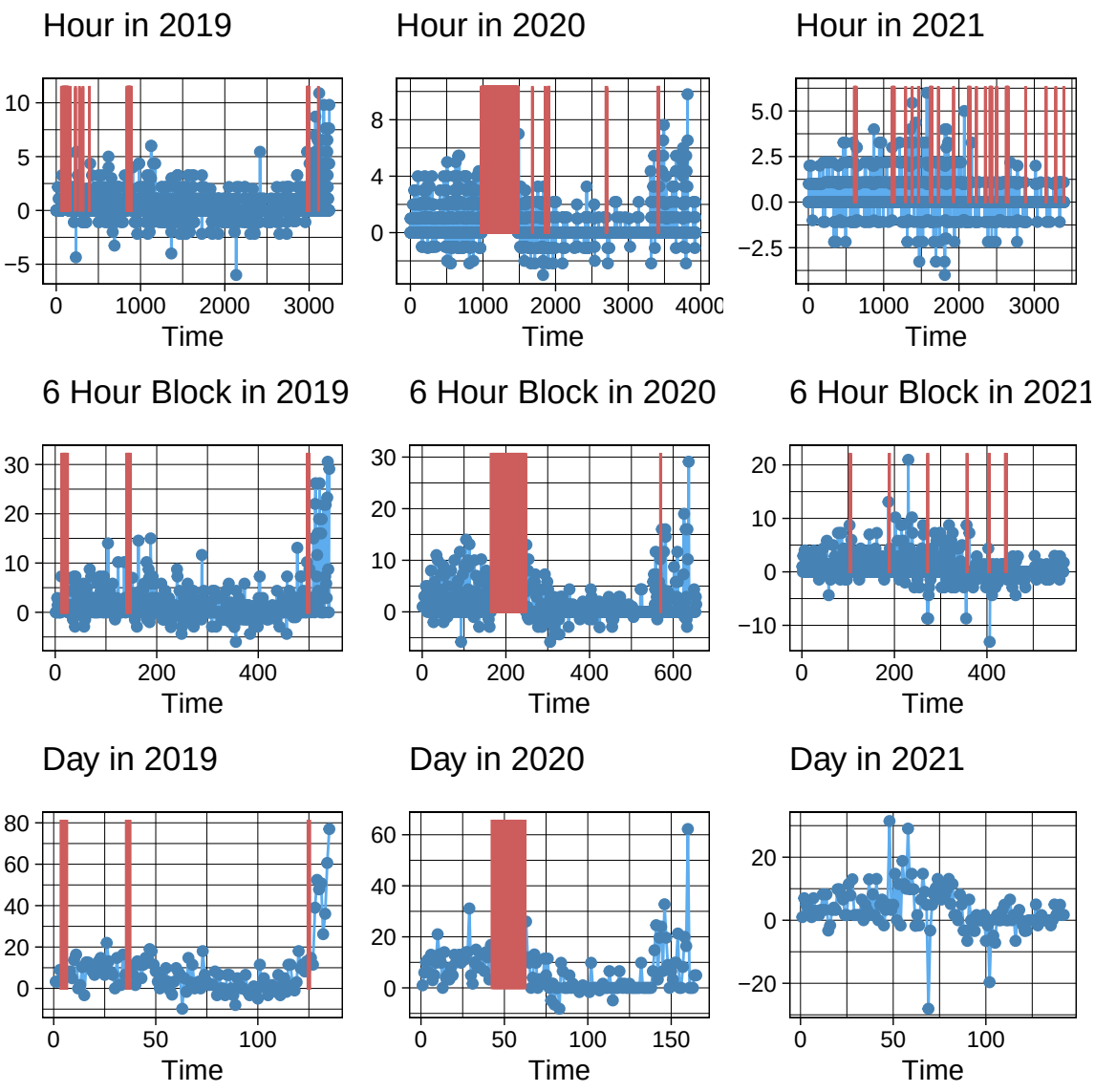

Figure 6: Time series of upstream fish in blue, with missing data highlighted in red.

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	64 (11)	158 (11)	138 (11)
6 Hour Block	2019	49 (17.6)	109 (17.6)	70 (17.6)
Day	2019	38 (19.2)	55 (19.2)	53 (19.2)
Hour	2020	50 (18.7)	17 (18.7)	10 (18.7)
6 Hour Block	2020	172 (28.1)	5 (28.1)	222 (28.1)
Day	2020	199 (35.5)	429 (35.5)	334 (35.5)
Hour	2021	3 (7.4)	3 (7.4)	0 (7.4)
6 Hour Block	2021	5 (7.7)	-2 (7.7)	2 (7.7)
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	738 (2.2)	802 (11.2)	896 (11.2)	875 (11.2)
6 Hour Block	2019	953 (5.2)	1002 (18.3)	1062 (18.3)	1022 (18.3)
Day	2019	1061 (10.3)	1100 (21.8)	1116 (21.8)	1114 (21.8)
Hour	2020	681 (1.7)	731 (18.7)	699 (18.7)	691 (18.7)
6 Hour Block	2020	802 (4)	974 (28.3)	807 (28.3)	1024 (28.3)
Day	2020	888 (7.8)	1086 (36.3)	1316 (36.3)	1222 (36.3)
Hour	2021	362 (1.5)	365 (7.6)	365 (7.6)	362 (7.6)
6 Hour Block	2021	452 (3.2)	457 (8.4)	450 (8.4)	454 (8.4)
Day	2021	482 (5.6)	-	-	-

4 Discussion Points

- What to do with rows where `data_recorded` is "Partial?" This includes 216 rows, or 1.6% 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.

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 57.9 cm is excluded from the steelhead counts.

- Incorporating some kind of spline or curve to account for steelhead run timing could improve this model. For example a 55 cm long fish might be considered a non-steelhead early or late in the season, but could be predicted to be a steelhead in the middle of the run, because it may be more likely that any fish is a steelhead then. To fit this model, I'll need the date when each fish's length was taken as part of the species composition data.

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.