CBASS Data cleaning

Katie Lankowicz

2023-06-13

Casco Bay Aquatic Systems Survey

The CBASS project has been ongoing since 2014, with a seine component in the greater Portland region of Casco Bay every year. The purpose of this document is to create a protocol for data cleaning and combination for further analysis. This is in anticipation of incorporation of data from the recently-started Harpswell CBASS component.

The data for 2014 - 2021 (excluding 2019) are stored in the Excel workbook `raw-seine-data.xlsx` across the sheets named `sites`, `species`, `trips`, and `fish`. The data for 2022 are stored in the Excel workbook `2022_Raw_Seine_Data.xlsx` across sheets of the same names. The data for 2019 are currently missing. The physical datasheets need to be located and entered into digital format. There are also suspiciously few sampling days recorded for 2018, so there may be yet-to-be-entered datasheets from 2018 to locate.

Data loading and combination

Data from Excel workbooks will be loaded in such a way that the individual sheets within the workbook will become dataframes within a list item in R. We will keep only the sheets mentioned above in cases where other data are provided. It's also important to ensure that variable names (column names) are standardized so that we can merge data from all years into a single dataframe later on.

Load and clean 2014-2021 data The bulk of the data are stored in this Excel sheet. There are extra sheets to remove, quality control issues to address, and site names to standardize. The first step is to load the data as a list of dataframes, one dataframe for each Excel sheet in the Excel workbook. Necessary dataframes will be saved, and extraneous (blank) columns in these dataframes will be removed.

The sites dataframe needs to be cleaned to ensure sites are referred to in the same way across the years of the dataset. We will also remove freshwater sites at Highland Lake and any samples that only occurred at a specific site once.

The trip information also needs to be cleaned. Any trips that occurred at our removed sites need to be removed from the trip info dataframe. Variables will also be checked for validity. For example, one temperature is recorded as 147 degrees C. This clearly is inaccurate, and will be replaced with NA. Categorical variables will be forced to set levels; there are instances in which people wrote editorialized versions of what they were supposed to, and we do not need these extra details for our quantitative analysis.

Moving on, the biological data for the subsampled 25 individuals per species need to be cleaned. Species names will be checked for spelling errors. The information will then be linked to the new trip_id structure so fish can be accurately assigned to a physical location and time.

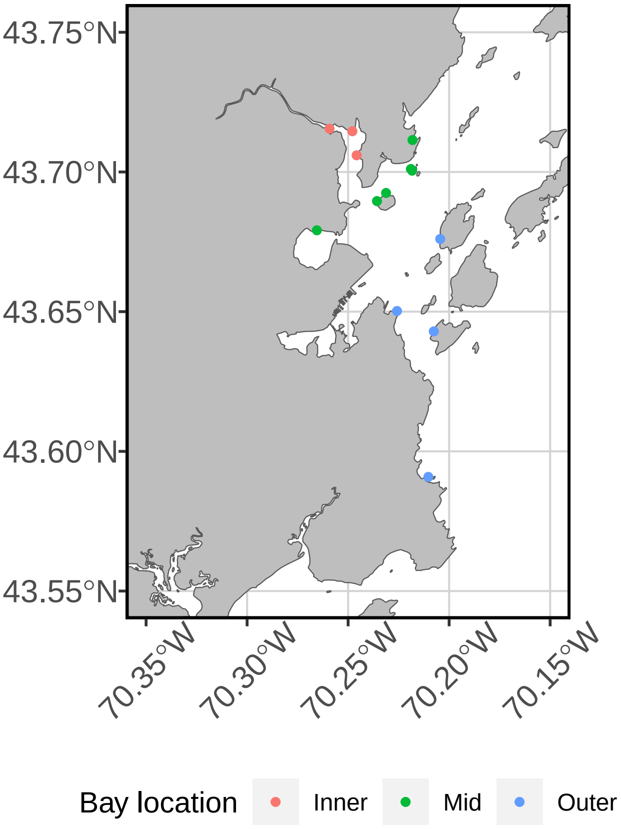
The abundance data need to be cleaned in a similar manner to the biological data.

Finally, the data will be QA-QC'd at a high level to ensure all variables are the correct format.

Load and clean 2022 data

Site information

The `sites` tab includes site name and geospatial location information (lat-lon, UTM northing-easting) for each sampling location. There are 13 potential sites in the greater Portland area of the survey. These are grouped into categories based on proximity to open ocean- “Inner Bay” being within the Presumpscot River region, “Mid-Bay” being north of Portland, and “Outer Bay” being closest to open ocean. We will plot these locations.



Data exploration

We will take the clean dataset and do some quick analysis and visualizations. This is always an important step for the start of a project; it helps to make sure your cleaning process went as planned, there are no obvious data gaps that will hinder your progress, and can even showcase any clear trends that may make for an interesting story.

Most-encountered species

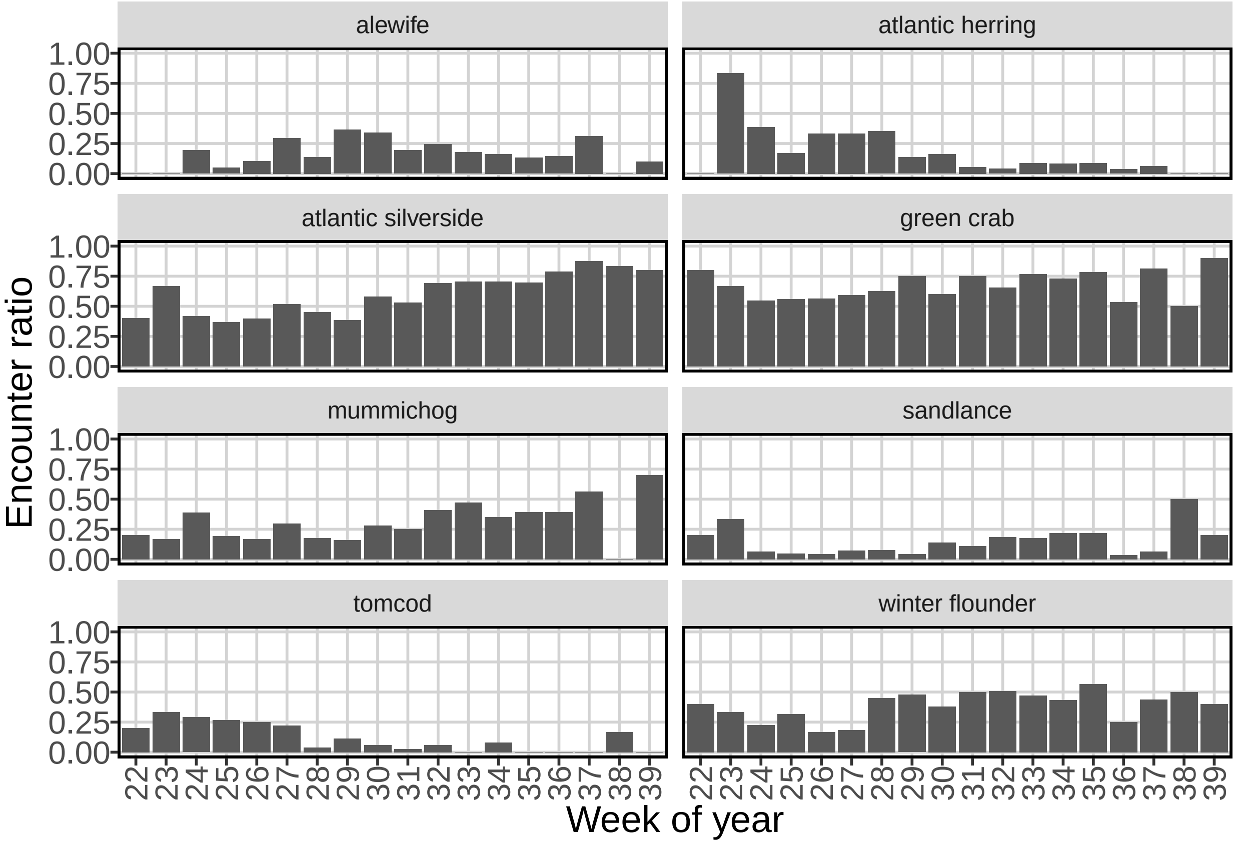
We will start by determining the top 8 species with the highest encounter ratio– number of encounters divided by total number of net hauls. This metric gives us more information than just total abundance OR raw number of encounters. Raw number of encounters does not account for varying number of sampling events year-to-year. Total abundance does not take into account the behavioral aspect of schooling fish. We want to know what species we're most likely to pull up in the seine while accounting for these factors.

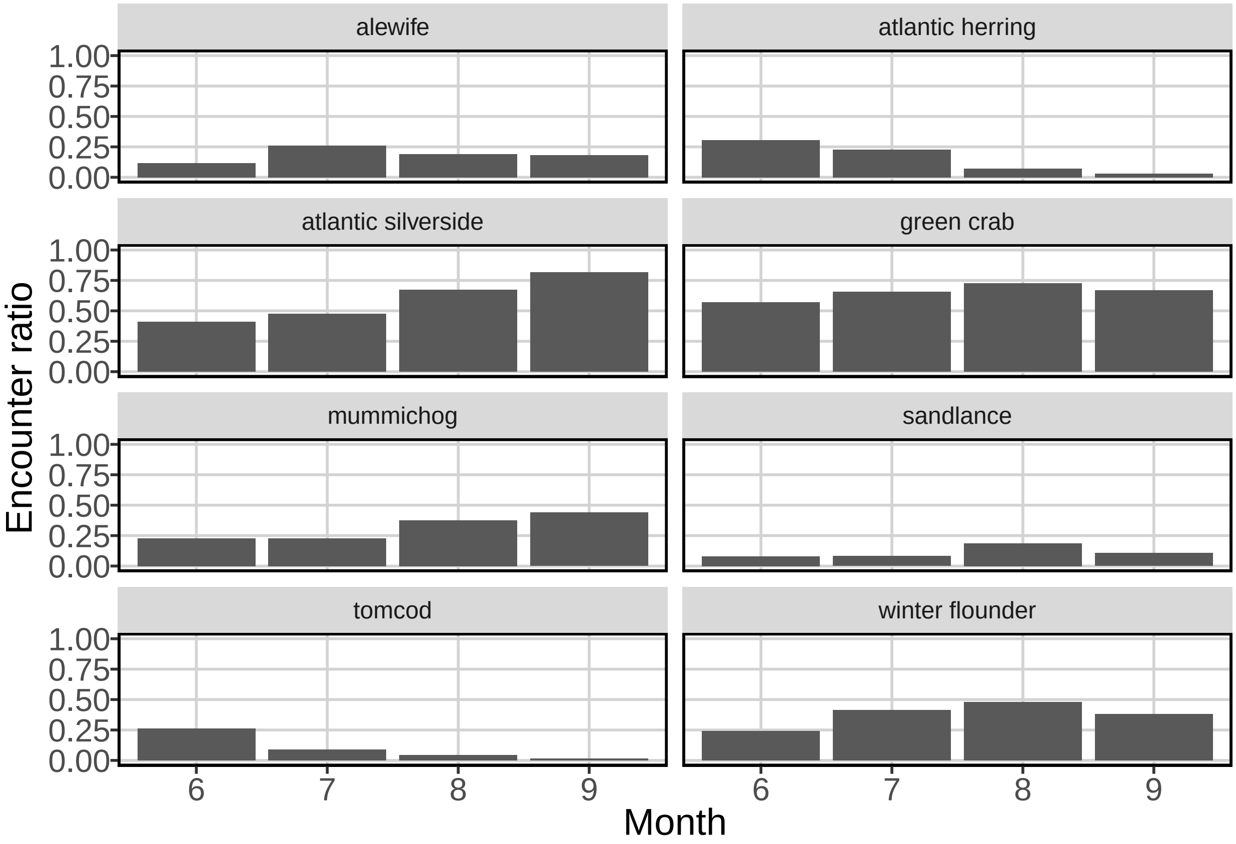
Table 1: Top 8 most abundant fish by encounter percentage

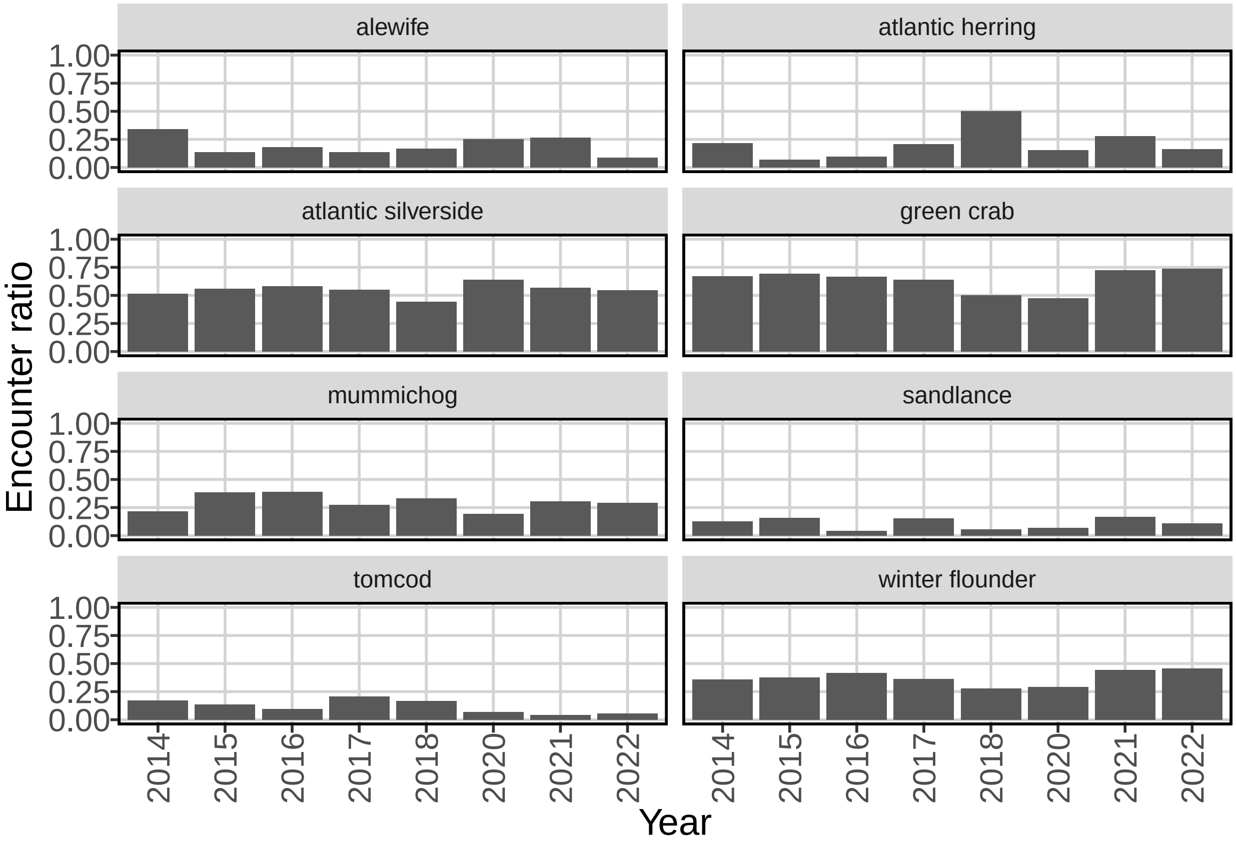
Species	Encounter ratio
green crab	65.7
atlantic silverside	56.1
winter flounder	38.6
mummichog	29.9
alewife	19.4
atlantic herring	17.5
sandlance	11.6
tomcod	10.9

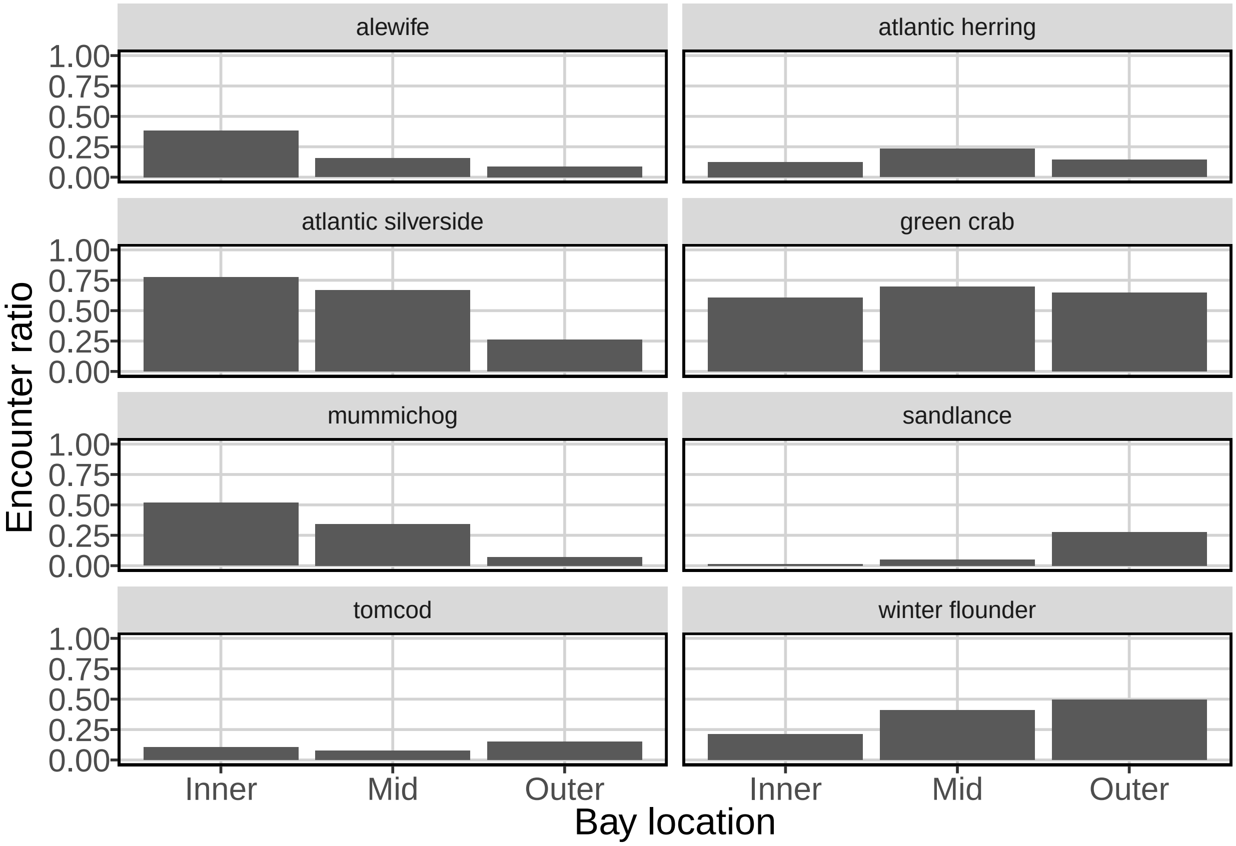
Encounter ratios

Next, we will plot the encounter ratios for every species across bay location, year, month, and week of year categories.



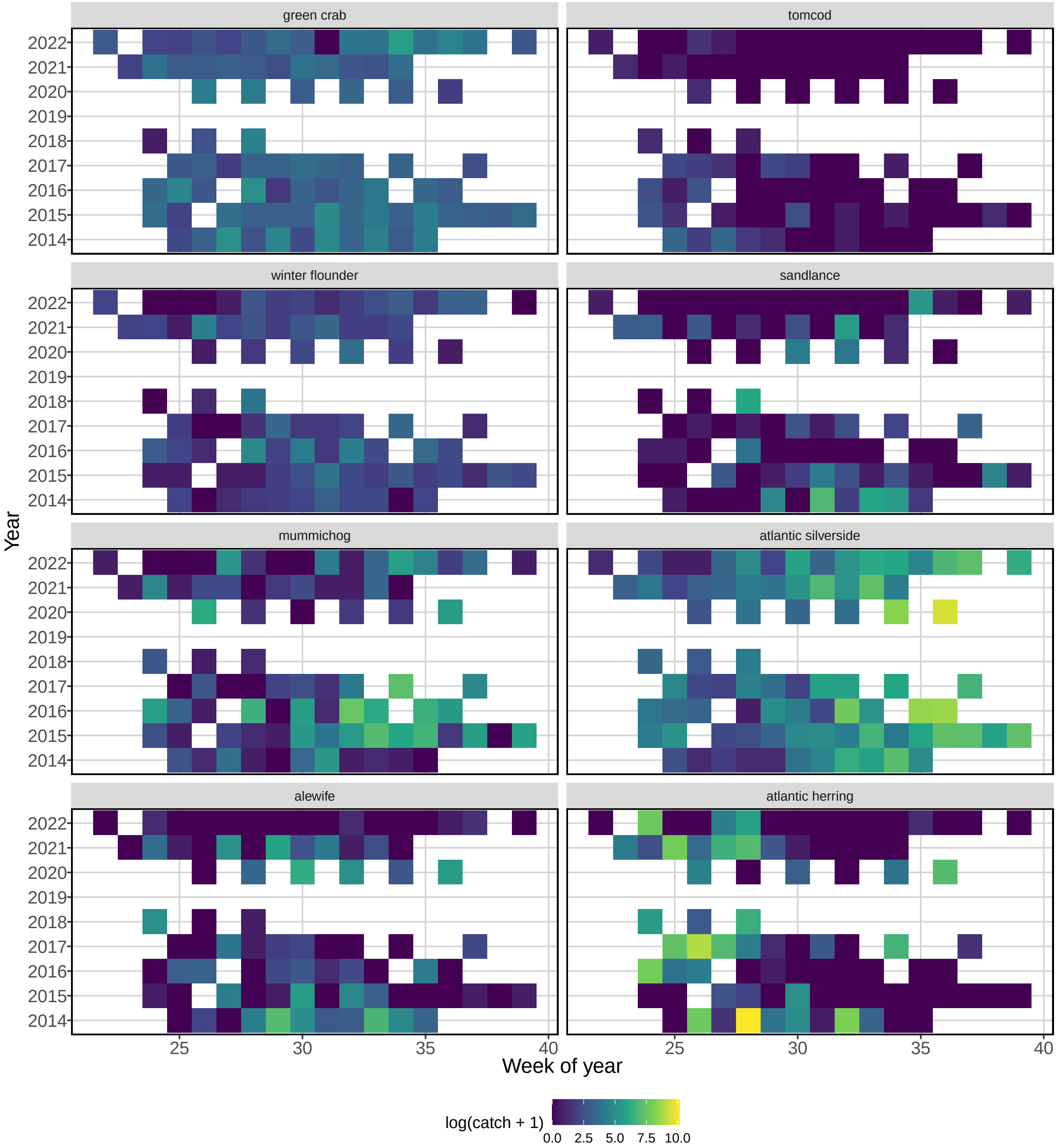

alewife
atlantic herring
atlantic silverside
green crab
mummichog
sandlance
tomcod
winter flounder
Encounter ratio
1.00
0.75
0.50
0.25
0.00
6
7
8
9
Month


alewife
atlantic herring
atlantic silverside
green crab
mummichog
sandlance
tomcod
winter flounder
Encounter ratio
1.00
0.75
0.50
0.25
0.00
2014
2015
2016
2017
2018
2020
2021
2022
Year


alewife
atlantic herring
atlantic silverside
green crab
mummichog
sandlance
tomcod
winter flounder
Encounter ratio
1.00
0.75
0.50
0.25
0.00
Inner
Mid
Outer
Bay location

Abundance periods

We can plot the abundance of each species over the survey period for each year. This will highlight any obvious trends in residence time.



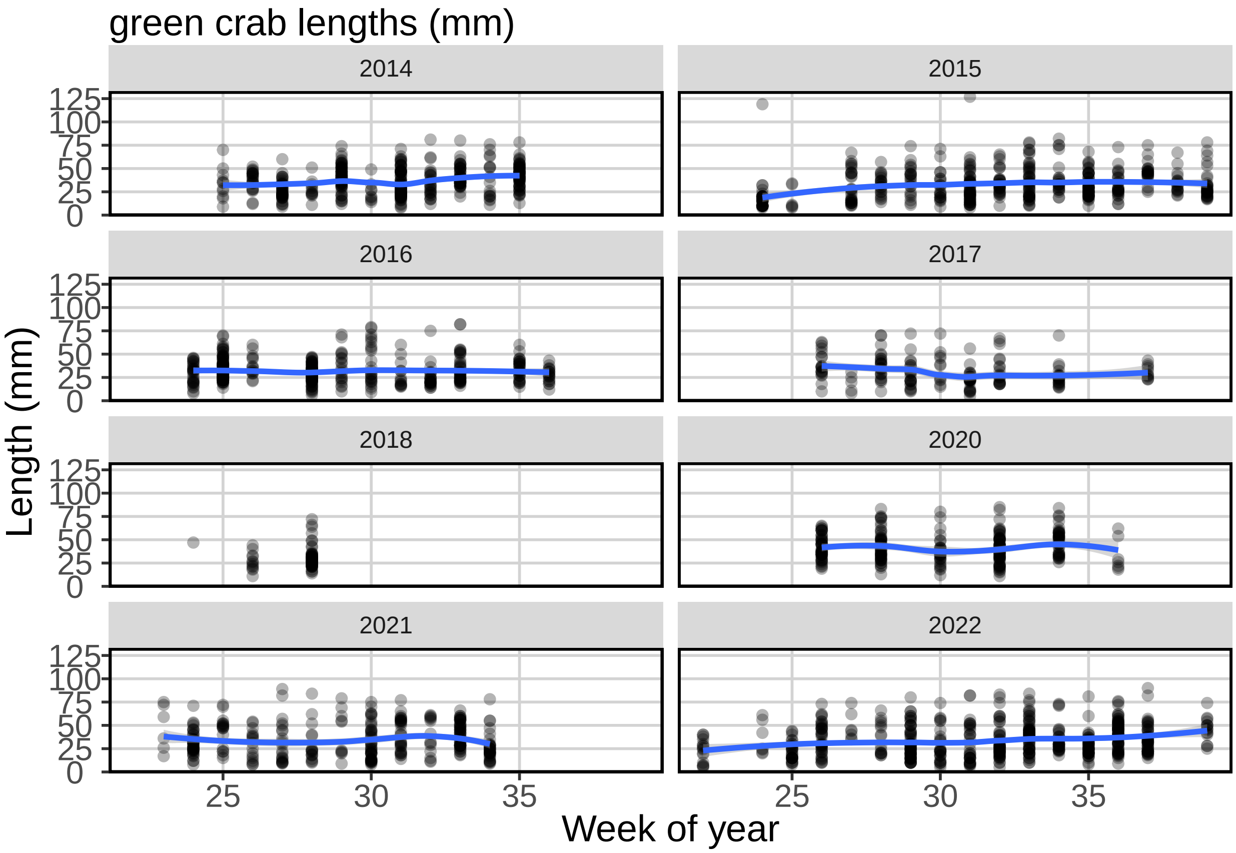
Size

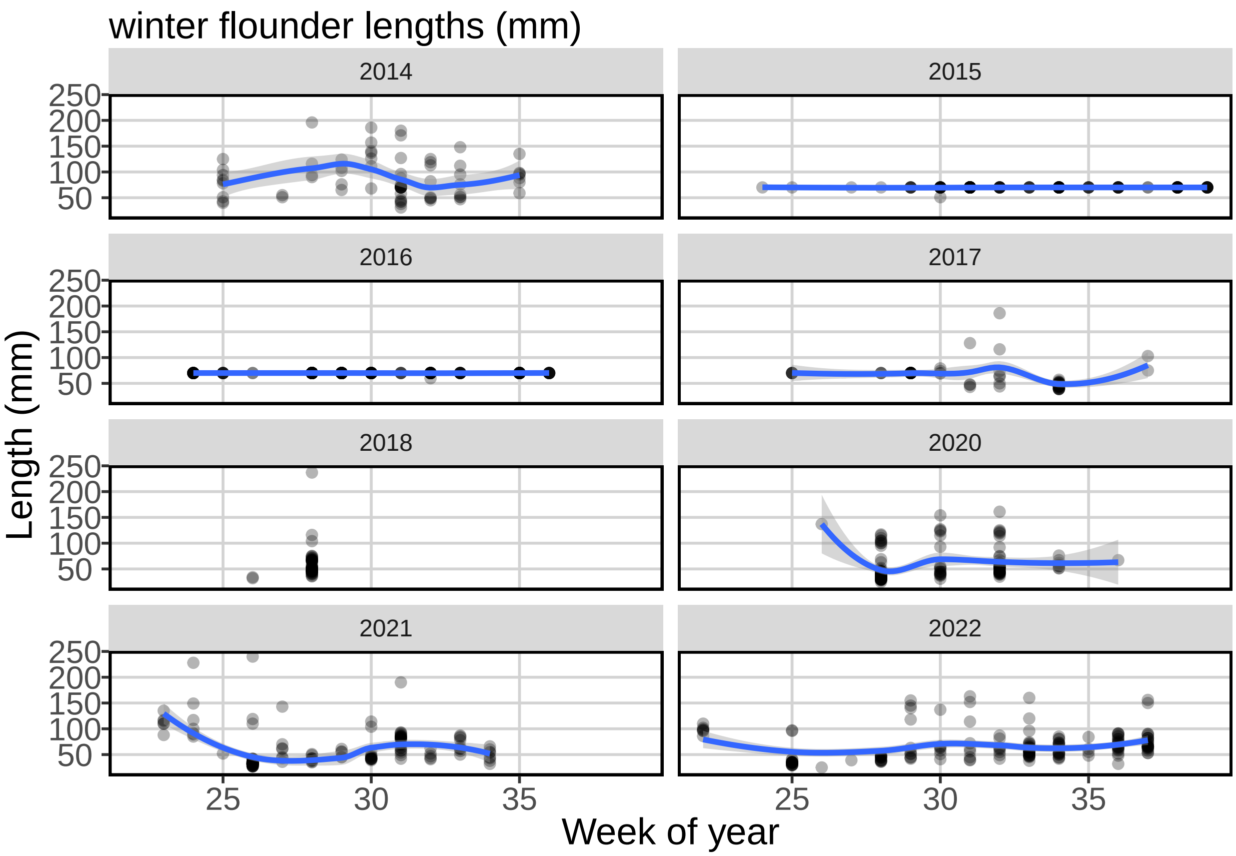
Next, we'll see if we can track growth of our most-encountered species by plotting lengths of sub-sampled individuals by week of the year.

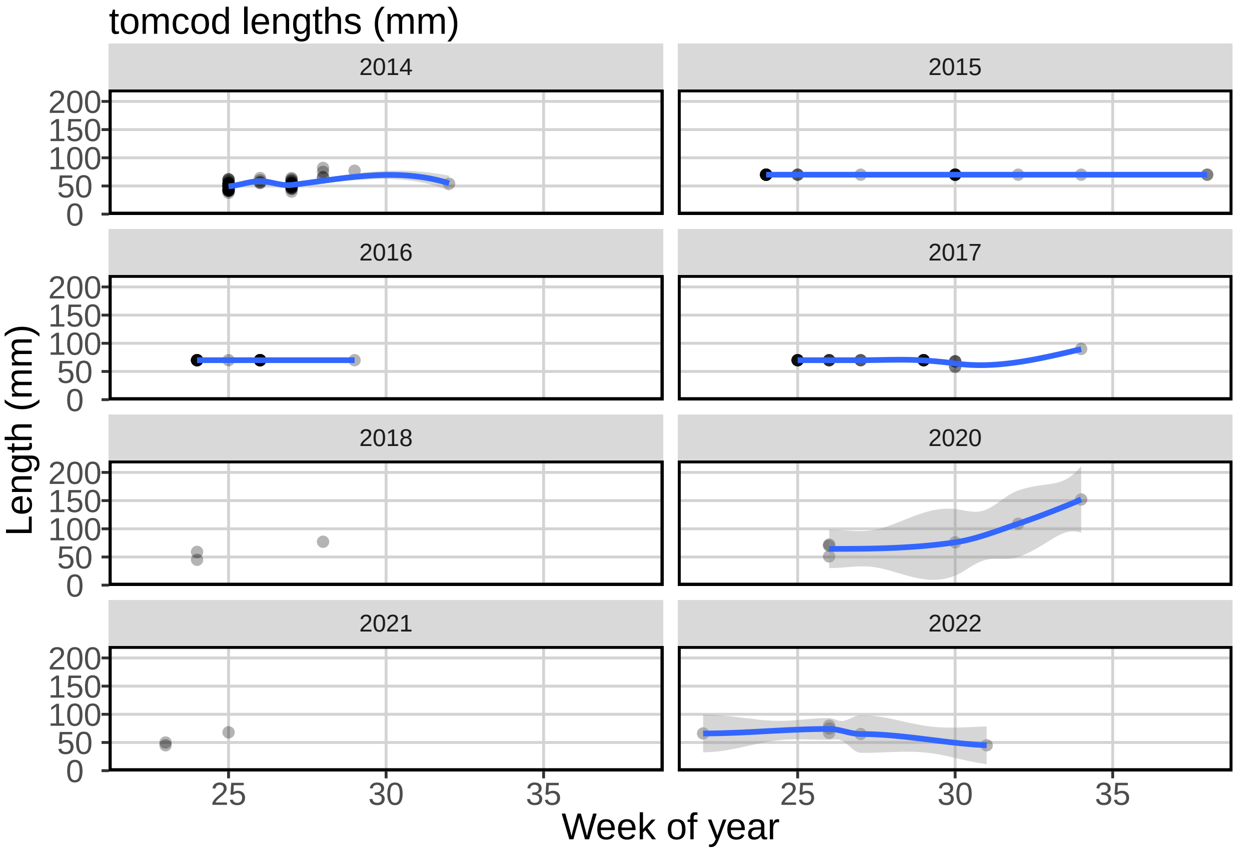
Honestly, the resulting plots look pretty terrible. It is likely that species with no clear growth over the summer period are constantly recruiting new individuals to the gear (spawning, and the babies take a few weeks to be big enough to be caught by the net). This looks to be the case for green crabs in particular.

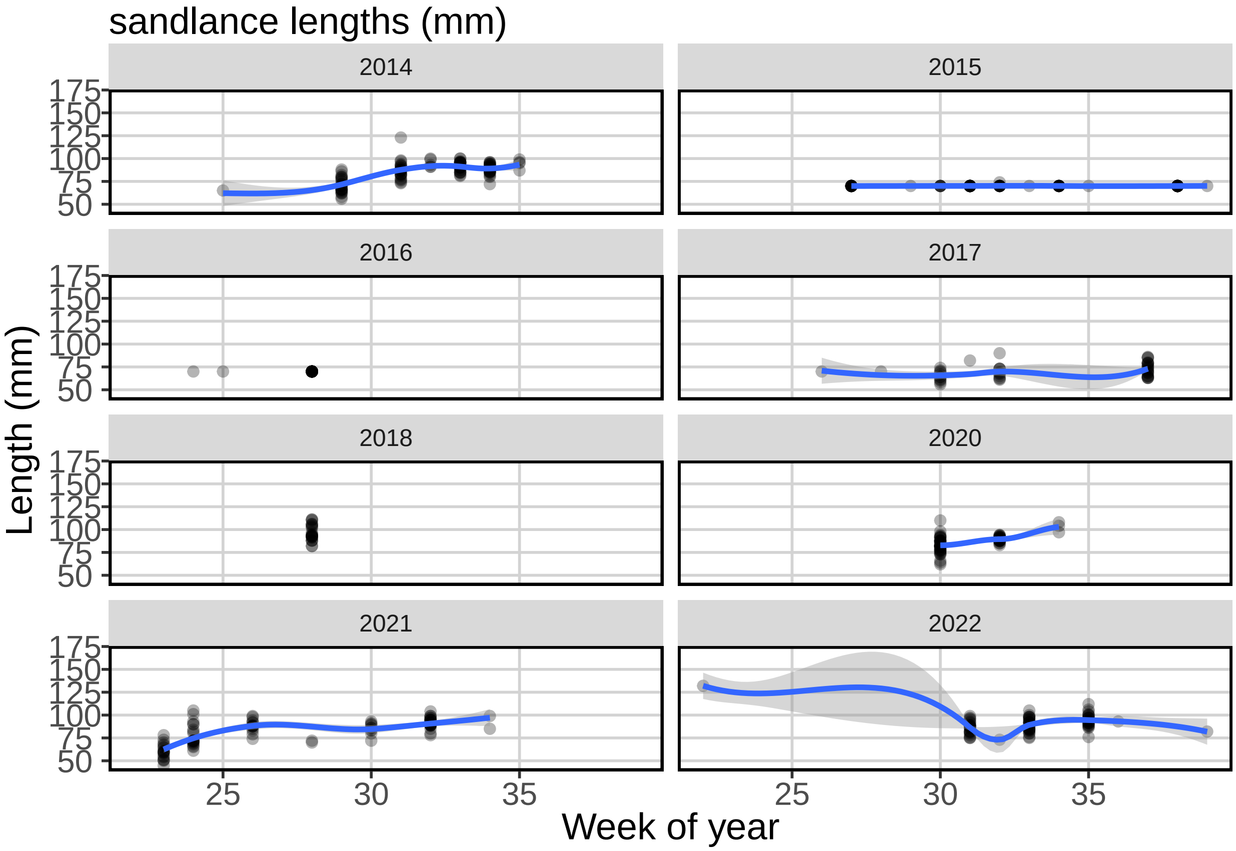
It's also possible that species with rapid declines in average length from early summer to mid summer are mixed cohorts– older individuals have overwintered in the nearshore area, are immediately available to the gear in June, and then it appears as though average size drops dramatically when the young-of-year cohort recruits to the gear in July. This is clearly what we see for alewife in a few years.

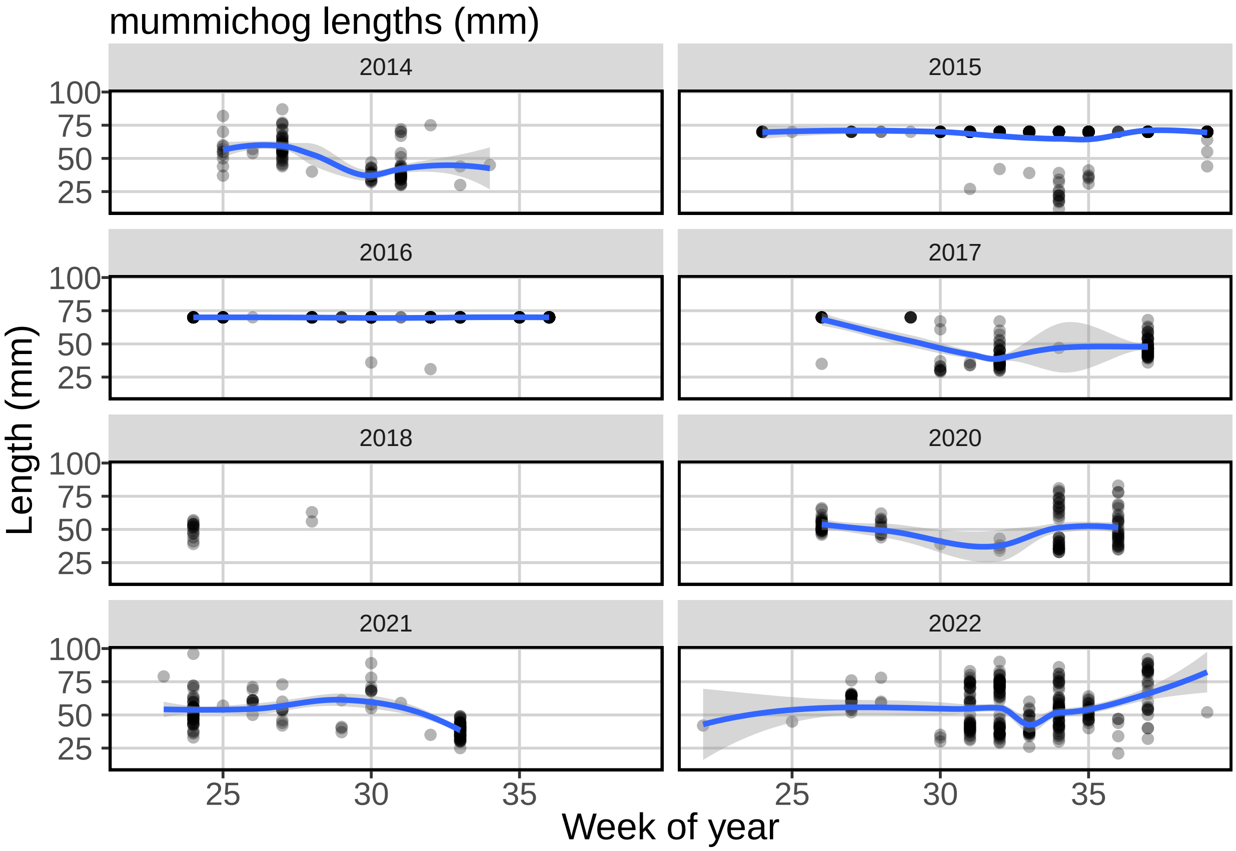
It would be beneficial to separate cohorts during analysis, if we want to look at growth rates. As is, we are still exploring and we'll keep them for now.

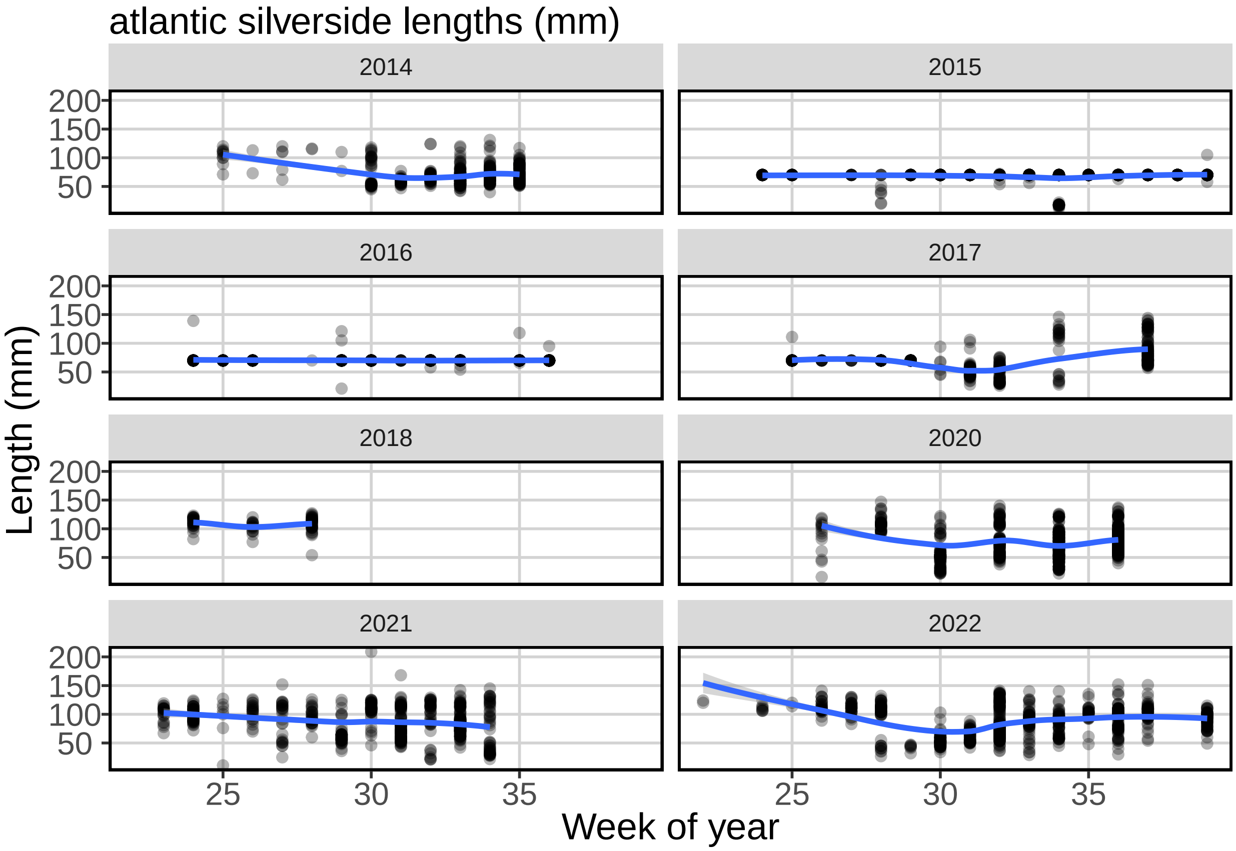


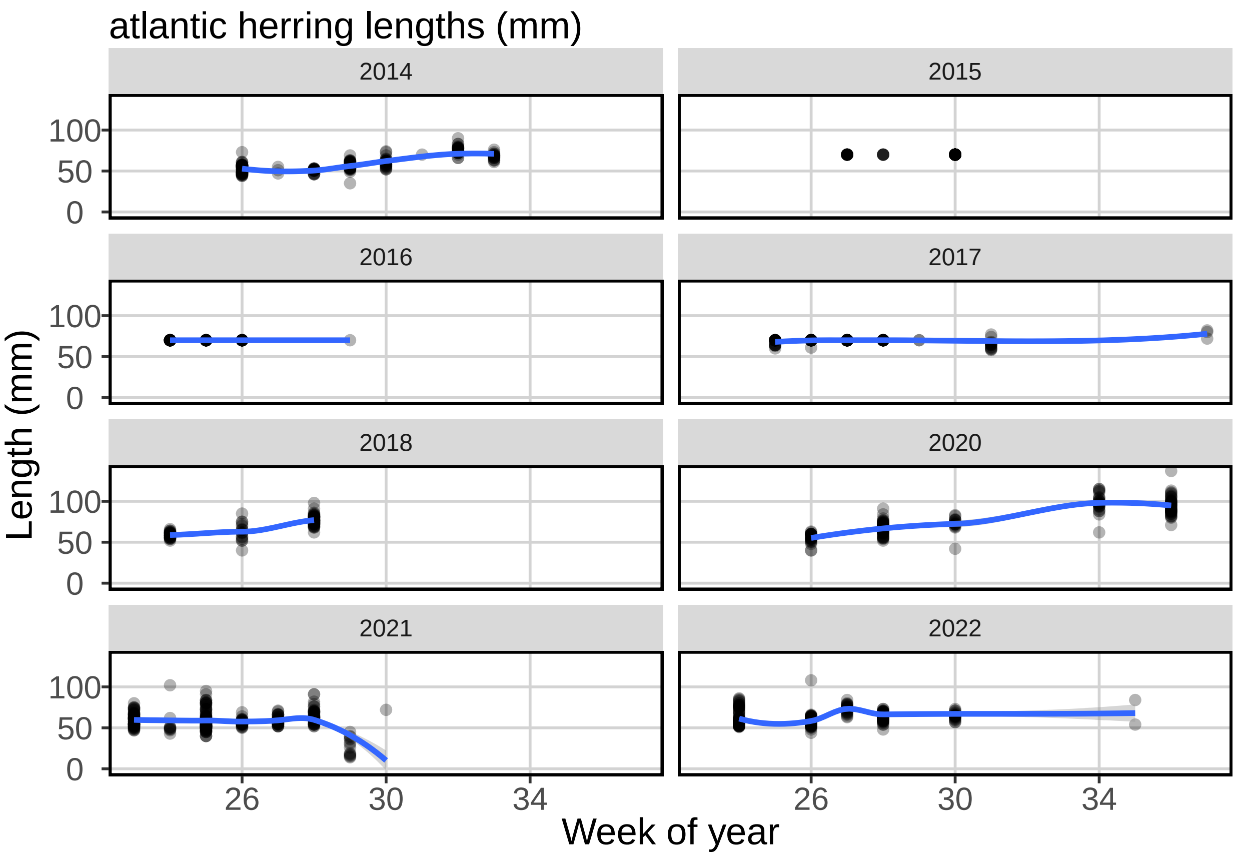

winter flounder lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
250
200
150
100
50
25
30
35
Week of year

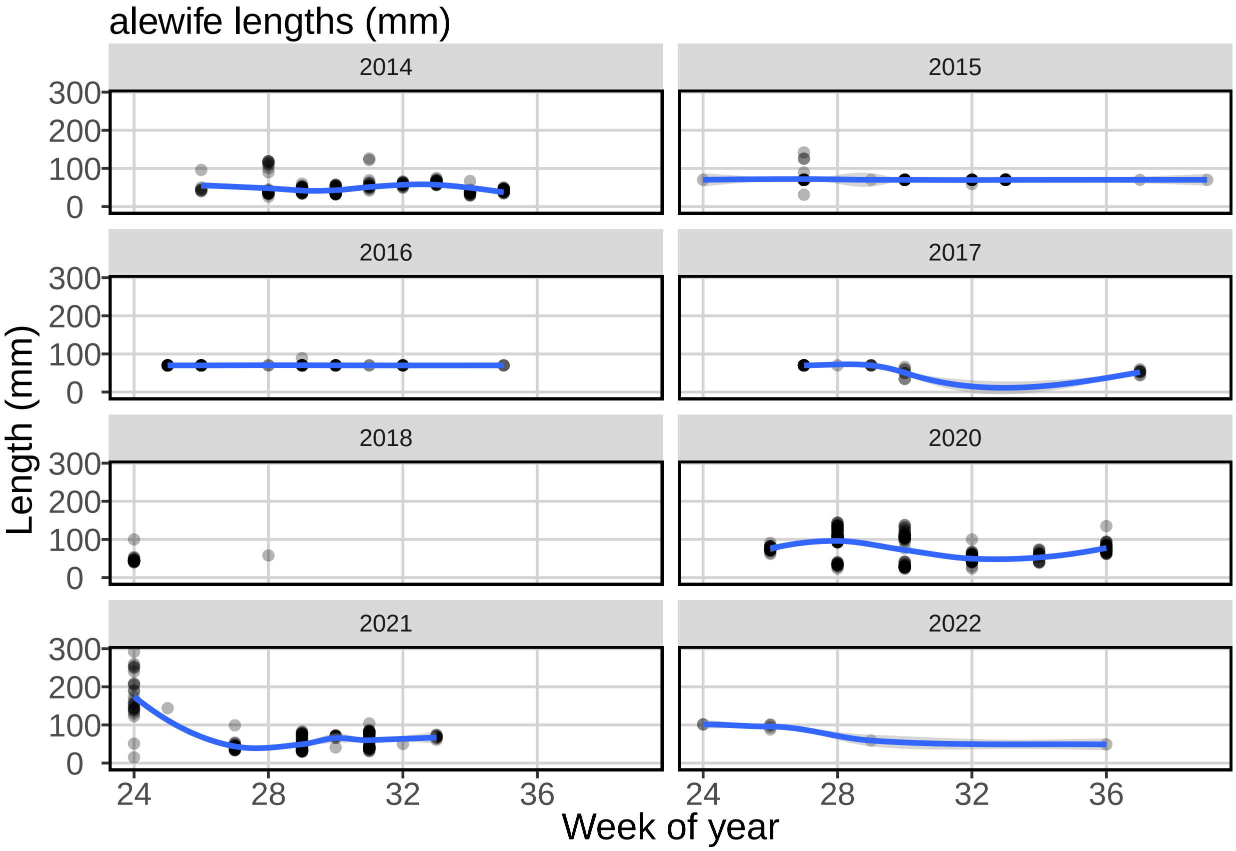

tomcod lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
200
150
100
50
0
25
30
35
Week of year


sandlance lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
175
150
125
100
75
50
25
30
35
Week of year


mummichog lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
100
75
50
25
25
30
35
Week of year


atlantic silverside lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
200
150
100
50
Length (mm)
25
30
35
Week of year


atlantic herring lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
0
50
100
26
30
34
Week of year


alewife lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
300
200
100
0
24
28
32
36
Week of year

Basic correlations (take with a grain of salt)

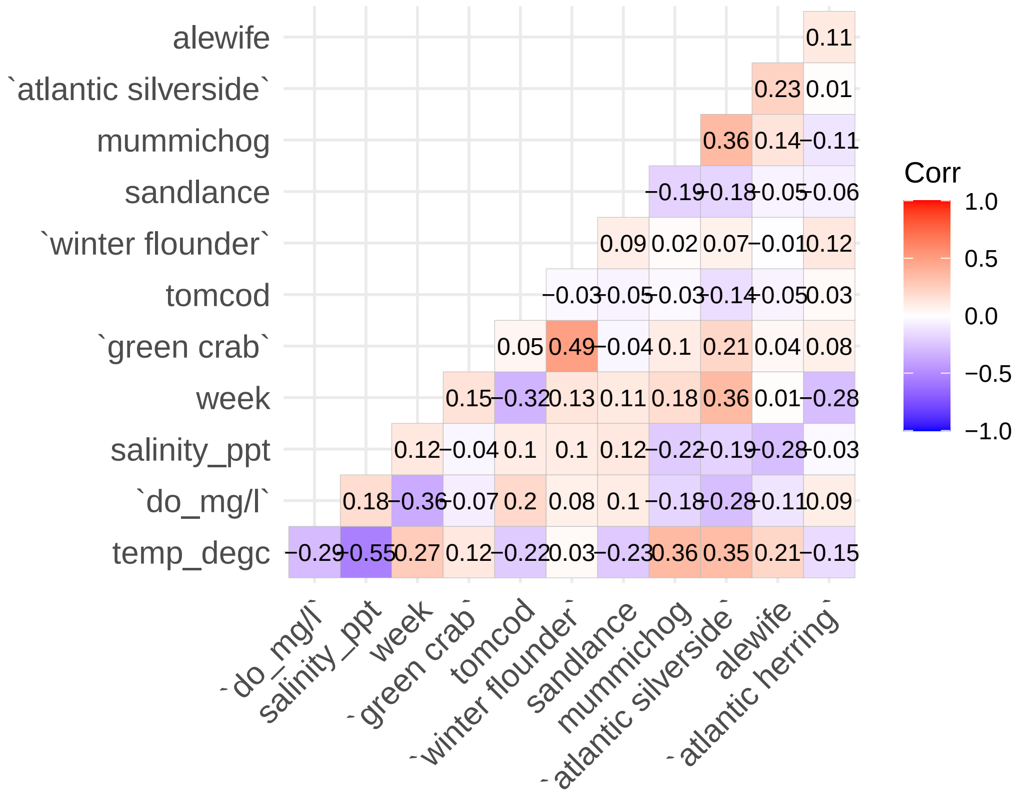


Figure 10: Density covariate correlation matrix