CBASS Data Cleaning and Exploration

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, then do some basic data exploration.

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 datasheets from 2018 to locate and enter.

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 sites that were sampled fewer than 10 times across the 8 years of the survey.

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 a typo, and will be replaced with 14.7 degrees C. 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 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 to ensure all variables are the correct format. We will also ensure that the number of fish reported in the abundance dataframe makes sense given the number of fish in the biological information dataframe. Recall that only 25 fish from each species are subsampled for biological information at each site; frequently, we will report higher abundance than number of fish measured However, we should never have more fish measured than reported in the abundance information. If this happens, we will use the number of fish measured as the correct number caught.

Load and clean 2022 data Data only from 2022 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 site information does not need to be cleaned. We will copy the site dataframe from 2014-2021 to the 2022 dataframe, because the same sites were sampled in these two periods.

The trip information will be cleaned similar to the 2014-2021 dataset. Freshwater sites do not exist in this year, but there are spelling errors to address. The site numbers and loc_id variables will be merged to match the patterns that exist in the 2014-2021 dataset.

Biological subsample information will be cleaned, but it is generally correct and doesn't need much adjustment.

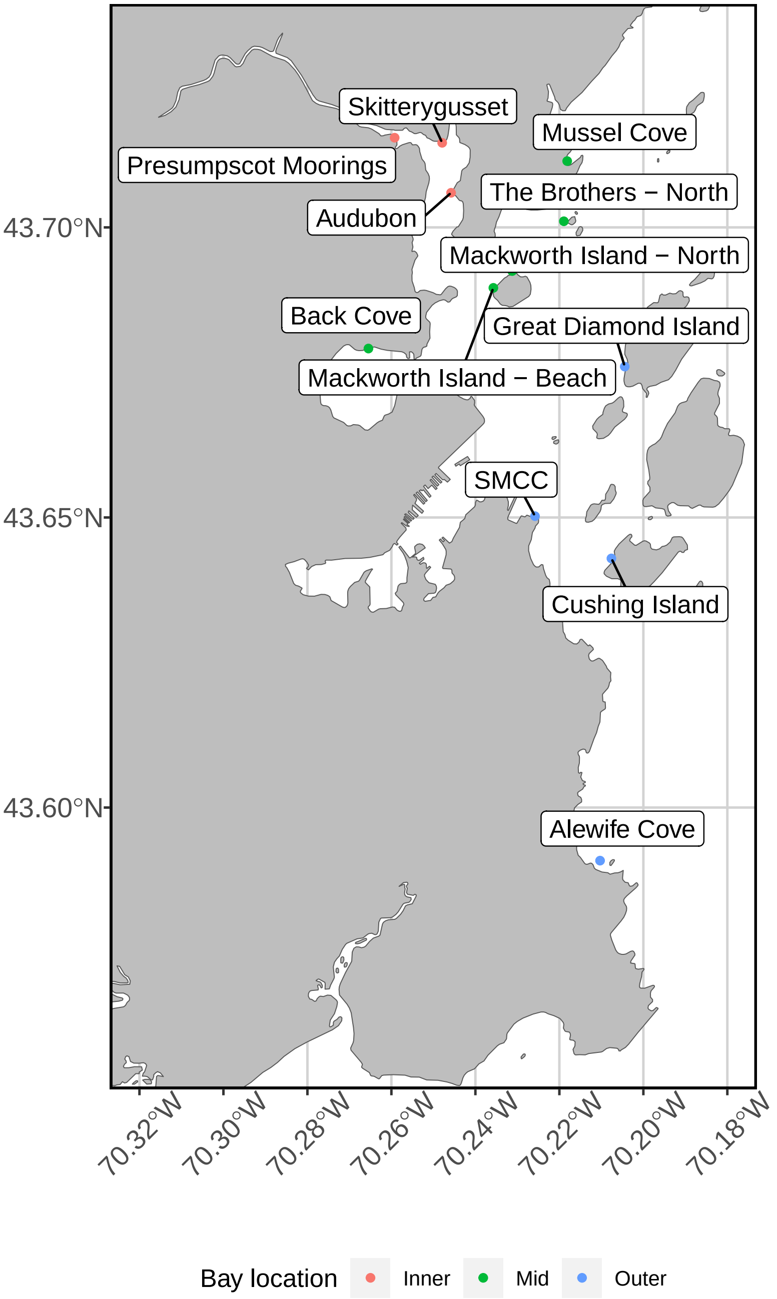
The same is true of the abundance information.

As a final check, we will ensure that species names are represented correctly between biological information and abundance datasets.

Combine data It's then a simple matter of combining the 2014-2021 and 2022 datasets to make a full record of this project.

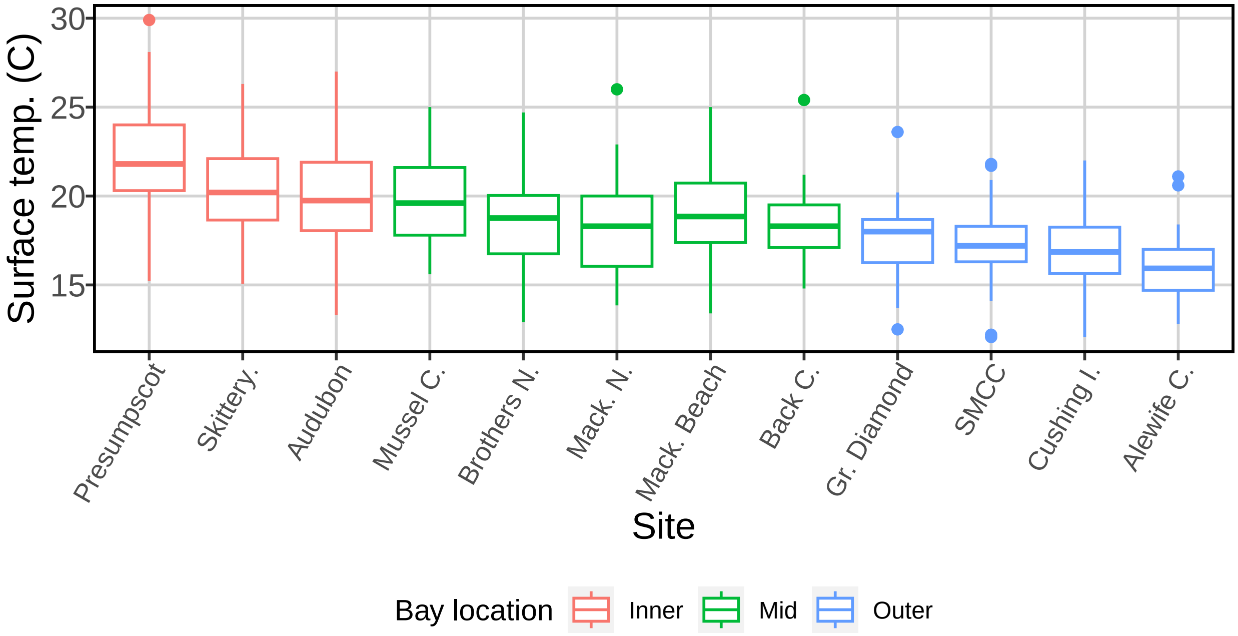
Site information

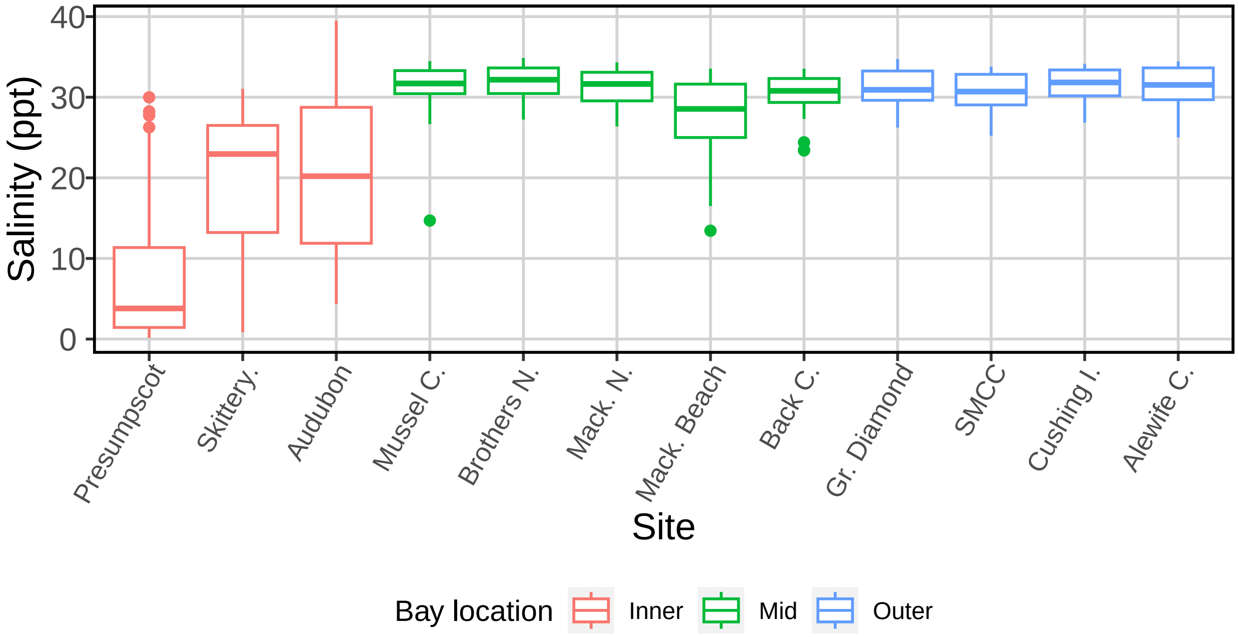
The sites tab includes site name and geospatial location information (lat-lon, UTM northing-easting) for each sampling location. There are 12 sites in the greater Portland area of the survey. These are grouped into levels of a category called “bay location” based on proximity to open ocean- “Inner Bay” being within the Presumpscot River region, “Mid-Bay” being north of Portland peninsula, and “Outer Bay” being closest to open ocean.

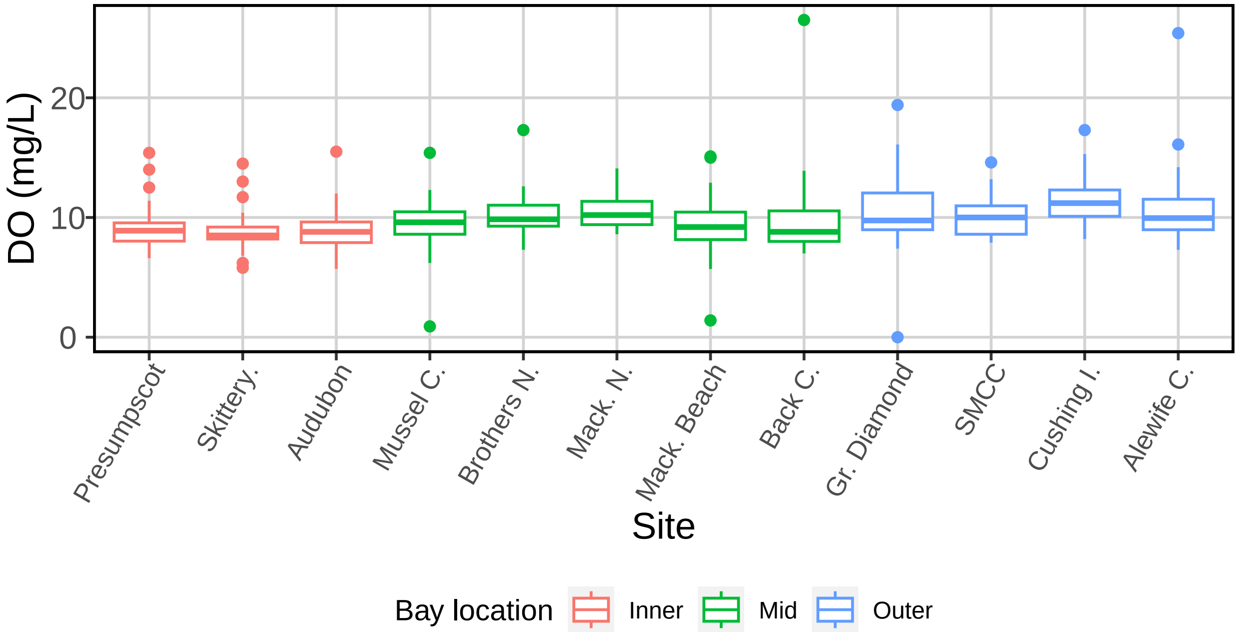


Environmental conditions

It is expected that sites within the same bay location level will have similar habitat and oceanographic qualities. We can quickly test this using our collected environmental variables – sea surface temperature, salinity, and dissolved oxygen concentration. Sites will be ordered by increasing distance to the head of the Presumpscot River.







These plots do not consider any temporal variation, but show on a broad scale that our bay location groups are fairly cohesive. There is a clear inner to outer temperature gradient, with sites inside the Presumpscot River being the warmest and those furthest from the mouth of the Presumpscot being the coldest. The inner bay sites have the lowest salinity. Middle and outer bay sites have similar salinity. Inner bay sites also have the lowest dissolved oxygen concentration; middle bay sites have slightly higher dissolved oxygen, and outer bay sites typically have the highest dissolved oxygen.

In the future, we can run statistical tests to quantify the similarity of the sites within our groupings. For now, we'll stick with this more qualitative assessment.

Data exploration

We will take the clean dataset and do some quick analysis and visualizations for the fish community structure and abundance.

What did we catch?

This is probably the most simple question we can answer. Without looking spatial or temporal shifts, we can report having caught the following species:

Table 1: Fish caught 2014-2022

Species	Total catch
atlantic herring	54949
atlantic silverside	52143
mummichog	12515
alewife	5136
sandlance	4653
green crab	3443
winter flounder	1226
atlantic menhaden	682
tomcod	182
bluefish	76
grubby sculpin	69
unID sculpin	58
shorthorn sculpin	46
northern pipefish	42
pollock	33
blueback herring	20
longhorn sculpin	18
rock gunnel	18
white sucker	18
threespine stickleback	16
banded killifish	15
mullet	14
smallmouth bass	14
atlantic cod	10
permit	9
american shad	8
unID sturgeon	7
white hake	7
horseshoe crab	6
striped bass	6
largemouth bass	5
smelt	5
butterfish	4
common dab	4
shortfin squid	4
unID shiner	4
lumpfish	3
ninespine stickleback	3
crevalle jack	2
emerald shiner	2
golden shiner	2
hake	2
northern puffer	2
white mullet	2
american eel	1
periwinkle	1
red hake	1
river herring	1
shortnose sturgeon	1
slimy sculpin	1
spotted hake	1
striped sculpin	1
unID gunnel	1

Most-encountered species

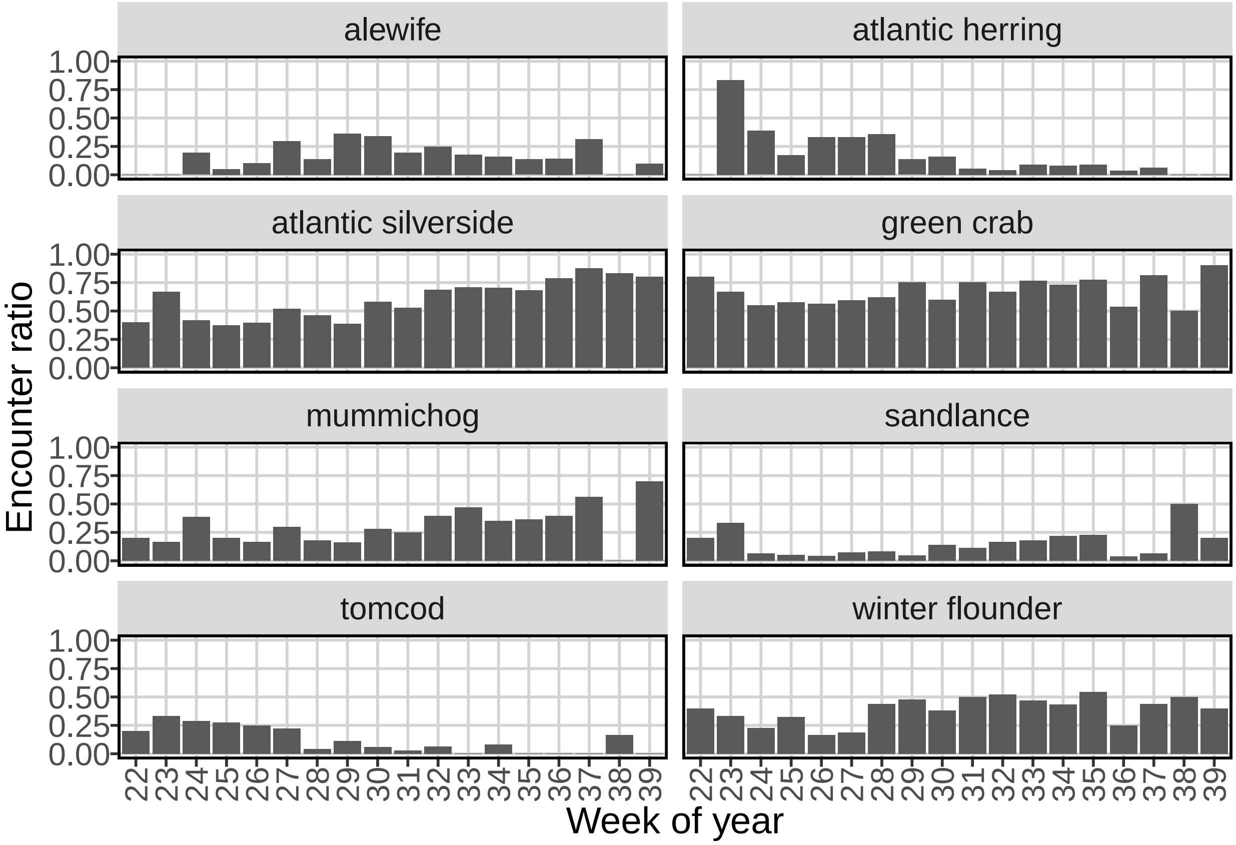
We will calculate encounter ratio for each species we have identified– number of encounters divided by total number of net hauls. This metric gives us more information than just total abundance OR raw number of encounters. Raw number of encounters does not account for varying sampling effort year-to-year. 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.

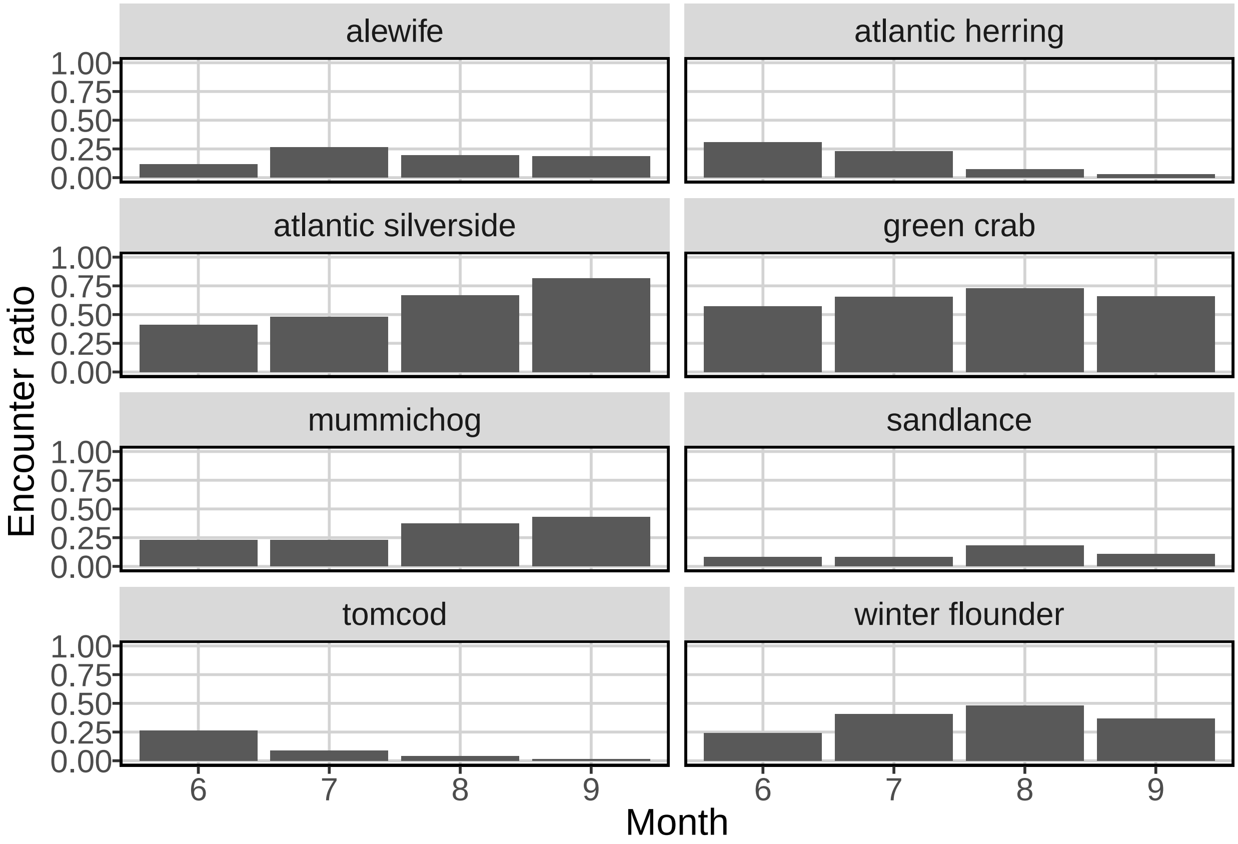
Table 2: Top 8 most abundant fish by encounter percentage

Species	Encounter ratio
green crab	65.8
atlantic silverside	56.1
winter flounder	38.5
mummichog	29.7
alewife	19.5
atlantic herring	17.7
sandlance	11.5
tomcod	11.0

Encounter ratios

We will plot the encounter ratios for the 8 most-encountered species across bay location, year, month, and week of year categories. Within the levels of each category (bay location, year, etc.), the encounter ratio will be calculated as the number of times that particular species was encountered divided by the total number of samples.





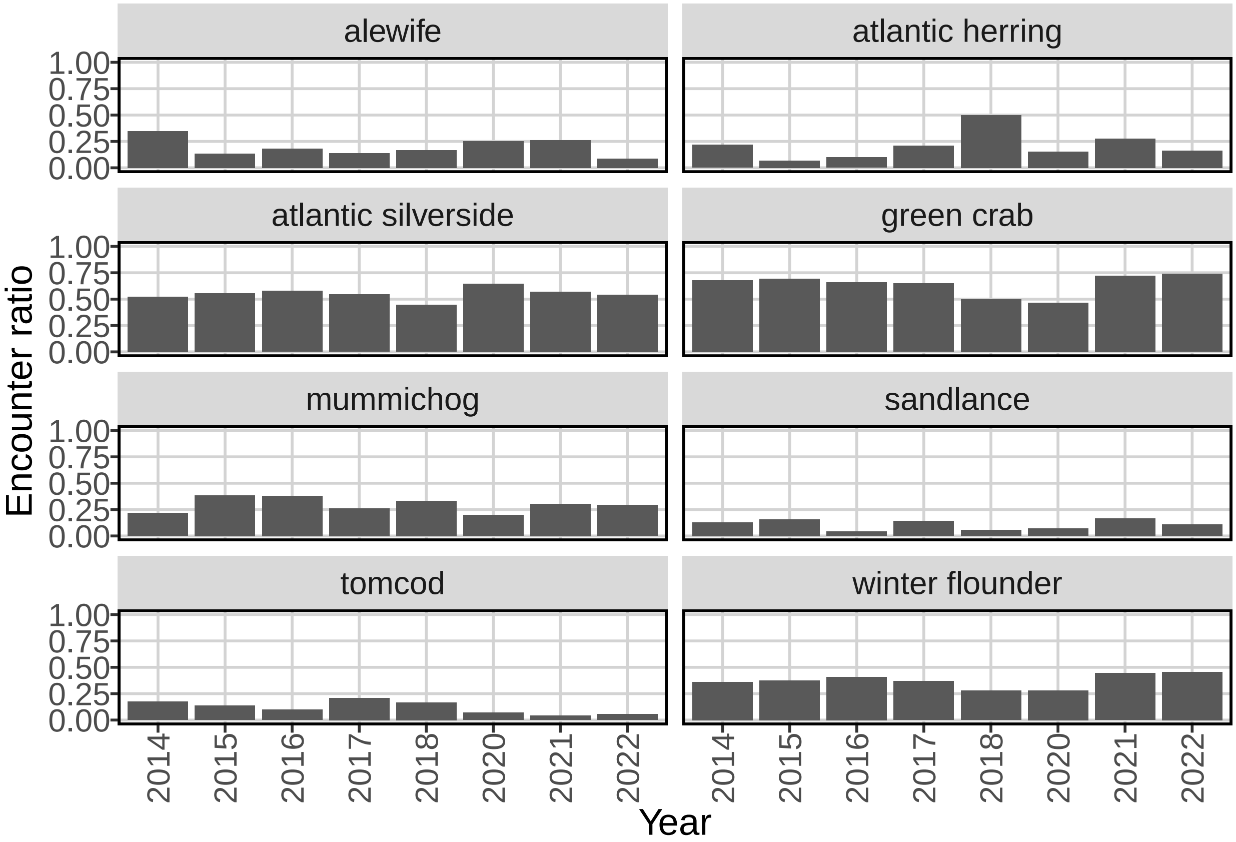
The weekly plot is good at showcasing trends in encounter frequency with a fine-scale temporal lens, but it may suffer from unaddressed issues with uneven sampling between years and temporal autocorrelation. The monthly plot, with a slightly more coarse temporal lens, smooths issues of uneven sampling but does not account for interannual variation or temporal autocorrelation. However, these plots still showcase interesting trends in phenology, residence time, and abundance.

Most obviously, Atlantic silversides and green crabs are highly abundant. They are sampled in at least 25% of net hauls in every week of the summer sampling period. Encounter ratio for silversides has an increasing trend from early June to early September, while the encounter ratio for green crabs is fairly stable and may only show a slight increase. Mummichogs, though not as commonly encountered, may also have a slight increase in encounter ratio through the summer. Atlantic herring and tomcod show the opposite trend, with decreasing encounter frequency through the summer.

The pattern of alewife encounters may reflect the downstream migration of fish spawned the previous year. It appears as though the early June - late August period encompasses most of this movement.

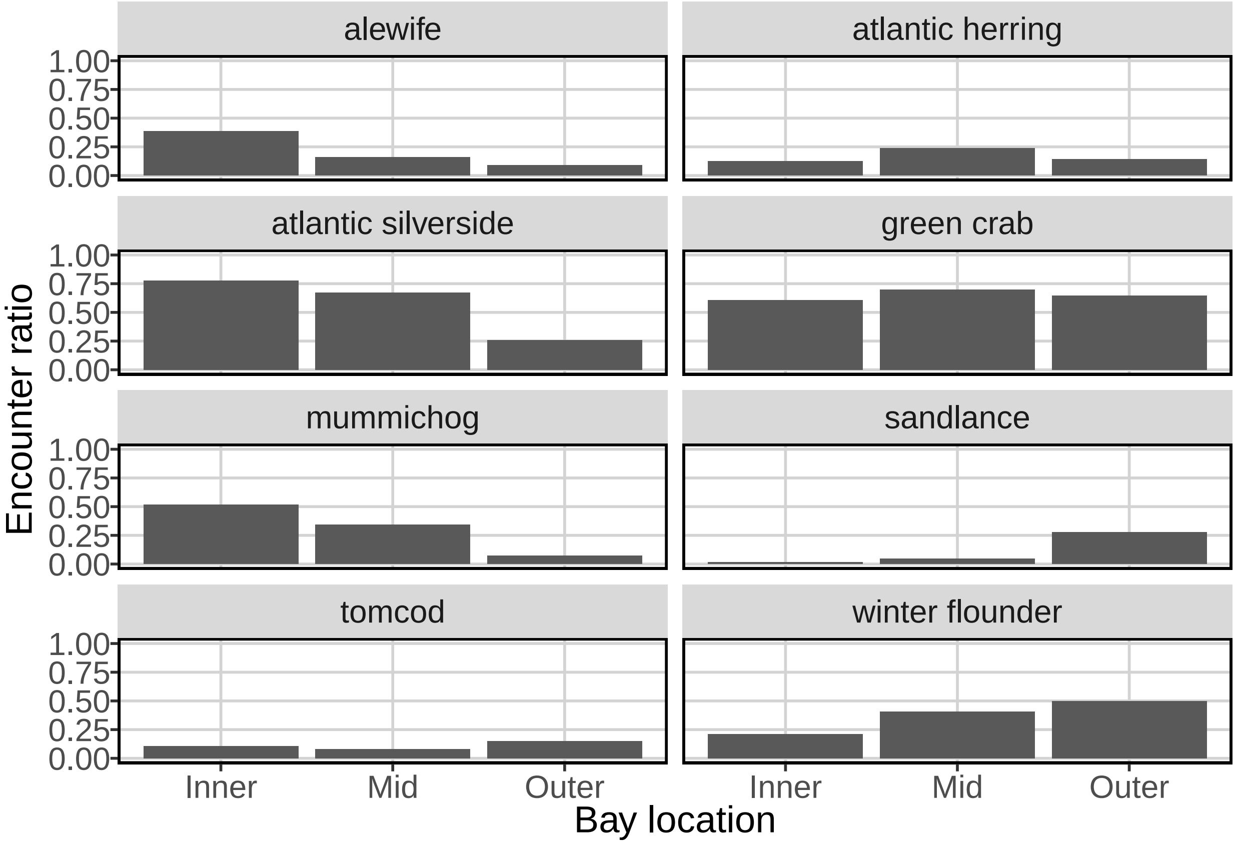
Winter flounder are frequently encountered, and appear to have slightly higher encounter ratios in the mid-late summer period.

Sandlance encounter ratios are highly variable week-to-week. It is likely that any patterns noted in their weekly or monthly encounter ratio plots are mostly driven by high interannual variability in their catch.



Trends in encounter ratios over the years give a quick overview of interannual variation. However, they may also reflect variable sampling effort– note that 2018 had very low sampling effort (sampling only occurred in 3 weeks across June and July) and encounter ratios for that year may be depressed or inflated in a way that does not reflect the true abundance of species in that year.

Most species have relatively stable encounter ratios during the studied years. The exceptions are herring and sandlance, which have high encounter ratio variance, and tomcod, which have decreasing encounter ratios from 2017 to 2022.



These plots illustrate likely habitat preferences of these species. Green crabs and herring may have slightly elevated encounter ratios in the middle bay region, but otherwise have very similar encounter ratios across the bay. Alewives, silversides, and mummichogs are most frequently encounter in the inner bay, with decreasing encounter ratios with increasing distance from this region. Sandlance and winter flounder have the opposite trend, with highest encounter ratios in the outer bay and lowest encounter ratios in the inner bay. Tomcod have similar encounter ratios across all regions, but may be encountered at a slightly higher rate in the outer region.

Abundance and residence time

We can plot the abundance of each species over the surveyed weeks for each year. This will highlight any obvious trends in residence time. We will use a raster approach, where every cell indicates the abundance of a particular species in a certain week of a year. Gray raster cells indicate that there was no catch for that species at any sampled site in that week. Clear raster cells indicate that there was no sampling conducted in that week. The remainder of the color scale indicates abundance, with low abundance being purple and high abundance being yellow. The first plot will force the same color scale across all species to also indicate relative abundance to each other. The second plot will force a unique color gradient for each species, which allows for closer examination of residence time and interannual abundance for each species.

A few things pop out here. Atlantic silversides have been identified in every sampling week of the survey. They are one of the most-abundant organisms, as evidenced by the predominance of warmer colors in the first plot. They seem to have seasonal recruitment to the gear due to either individual growth or movement into the inshore area, with more silversides caught in the later weeks of each year. Having pulled more than my fair share of very small silversides out of the seine mesh (which acts like a gillnet when they are that size), I'm inclined to say it's probably mostly individual growth. Their abundance relative to other organisms has not changed much over the years of the survey, though 2016 and 2020 had weeks with very high abundance.

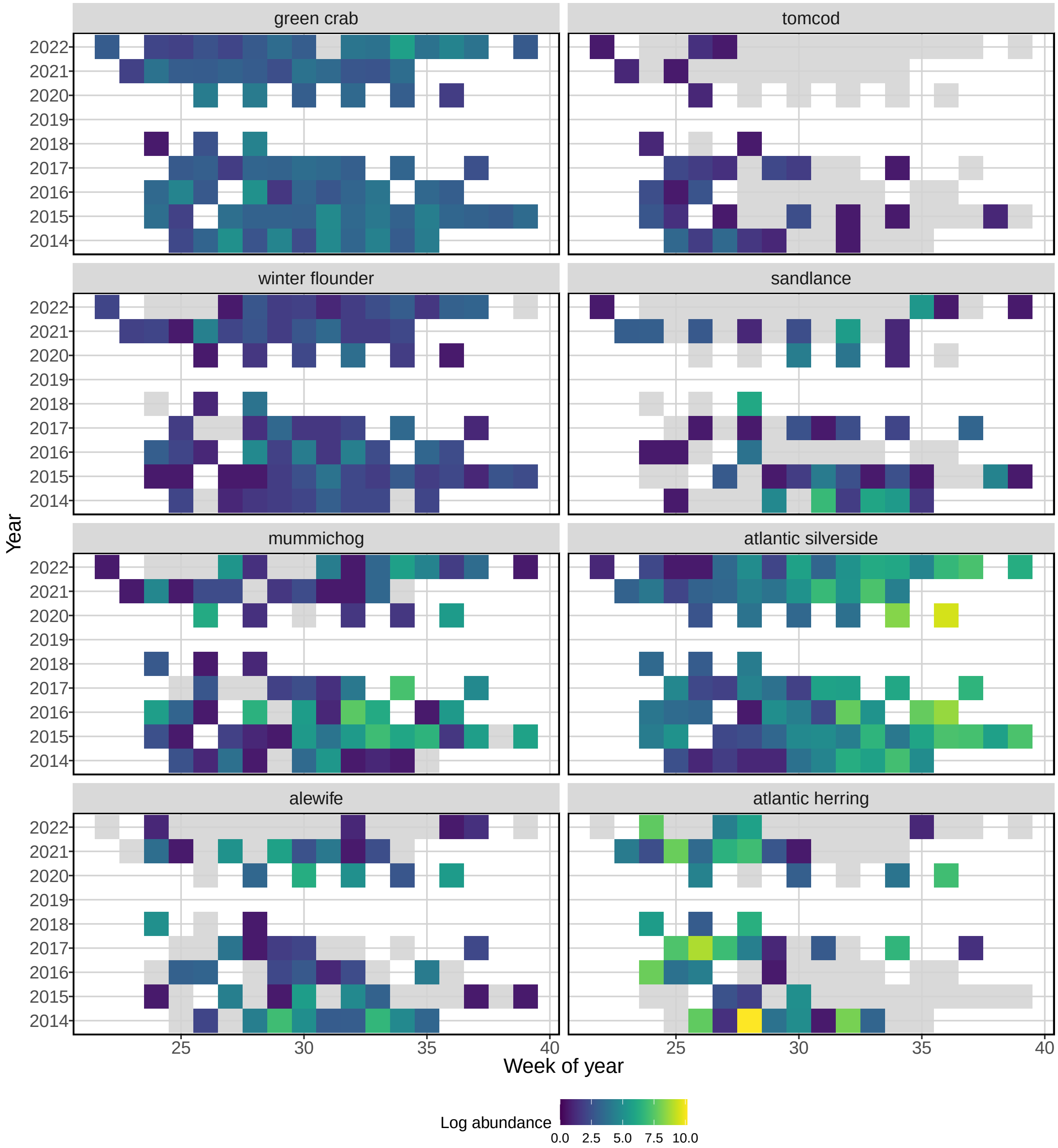
Only one week out of all sampling did not produce at least one green crab. Green crab abundance relative to other organisms has not changed much over the years of the survey. It also has not changed much seasonally.

Winter flounder are also seen fairly often. Though they do not have any obvious seasonal or interannual trends in relative abundance, their peak abundance occurred mid-summer in 2016.

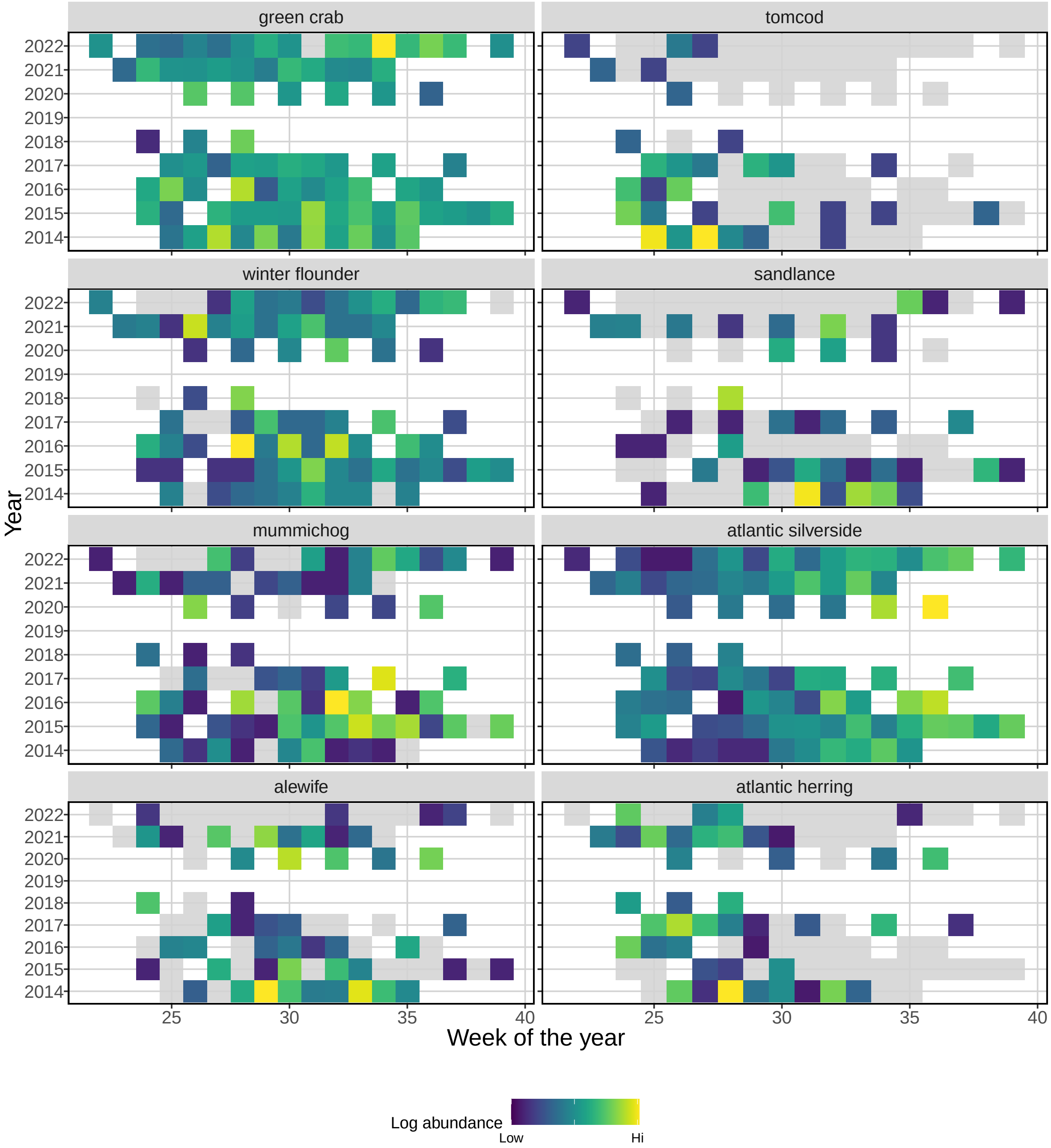
Mummichogs are caught more consistently and at higher abundances in the later weeks of the season. They have some inter-annual variation, with higher relative abundance in 2015-2016 than all other years.

Atlantic herring and tomcod have clear seasonal trends of abundance in the study area. For both species, abundance is highest in the earlier weeks of the season. Herring, which often form very large schools, have the highest single-week abundance of all 8 species, which occurred in mid-2014. Both herring and tomcod appear to leave the study area by late July.

There are no clear trends of abundance or residence time for sandlance and alewife. Both have high interannual variability and a peak abundance in mid-summer 2014.



green crab
tomcod
winter flounder
sandlance
mummichog
atlantic silverside
alewife
atlantic herring
Year
2022
2021
2020
2019
2018
2017
2016
2015
2014
Week of year
25
30
35
40
Log abundance
0.0
2.5
5.0
7.5
10.0

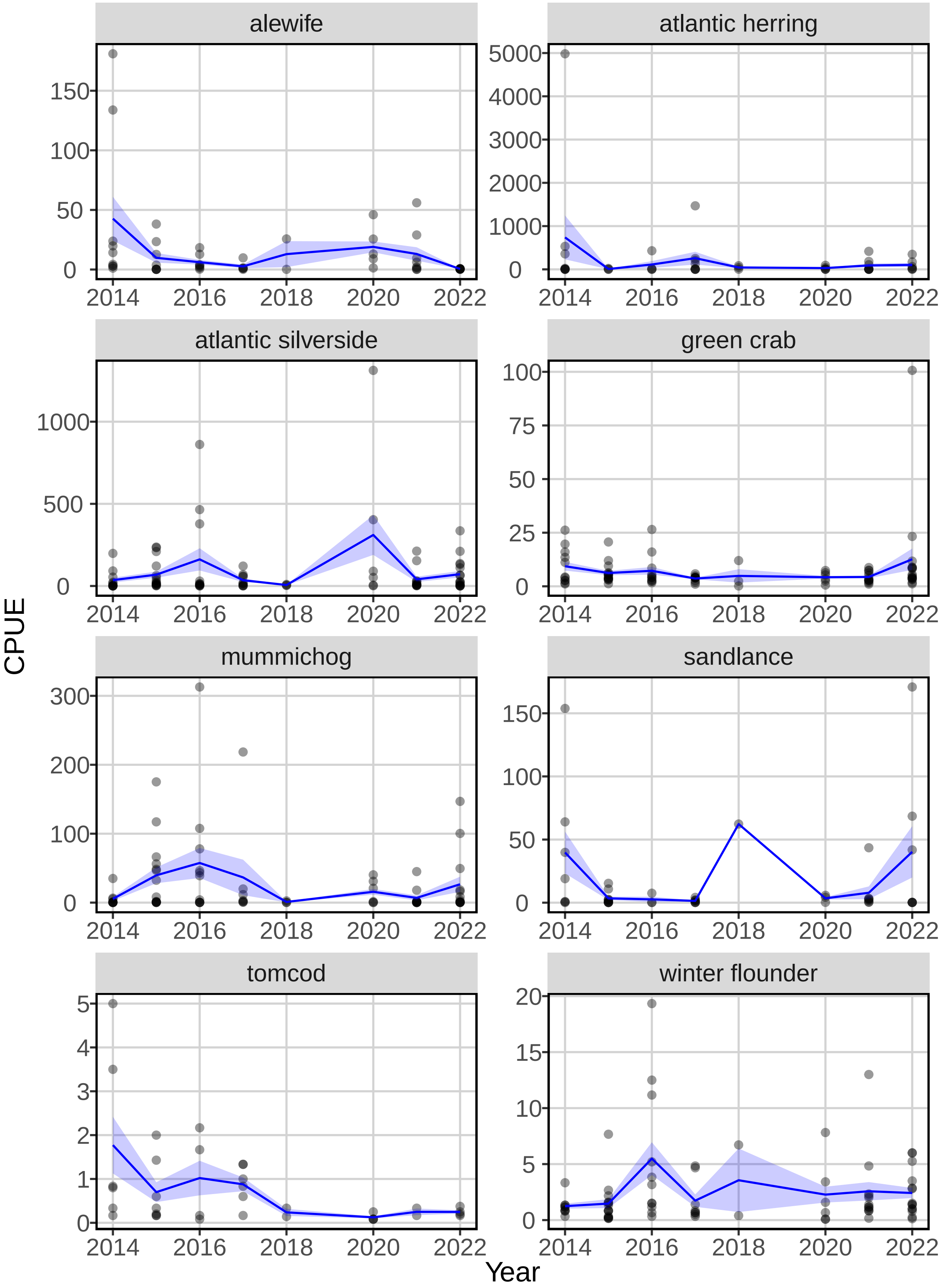

green crab
tomcod
winter flounder
sandlance
mummichog
atlantic silverside
alewife
atlantic herring
2022
2021
2020
2019
2018
2017
2016
2015
2014
Year
25
30
35
40
Week of the year
Log abundance
Low
Hi

Indices of abundance

This may be a foolhardy effort, but we can quickly generate some extremely basic and flawed annual indices of abundance for our 8 most-encountered species, then check out any trends. As the 2019 data is missing (but may exist), we will interpolate a linear trend between 2018 and 2020.

On these plots, the x-axis shows years of the survey and the y-axis shows catch per unit effort (number caught / number of net hauls). Points represent weekly measures of CPUE, while the blue line indicates mean CPUE and blue envelope indicates 95% confidence intervals. Please note that y-axis scales are unique to each plot. Indices for 2018 are not likely to represent true abundance trends, as most sites in this year were sampled only once or twice.

There are a few stories here that corroborate the abundance and residence time rasters above. Tomcod have a clear decline in abundance over time. Alewives and herring may also show a slight decline. Sandlance have very high interannual variability. Winter flounder and green crab populations have not changed much over time, though winter flounder had an exceptional year in 2016. Mummichog abundance is not quite as high 2020-2022 as it was 2015-2017, but does not have a clear trend. Atlantic silversides are typically one of the most abundant fish in the surveys, have some interannual variability, but look to be at about the same abundance index level in 2022 as earlier in the time series.


alewife
atlantic herring
atlantic silverside
green crab
mummichog
sandlance
tomcod
winter flounder
CPUE
Year
0
50
100
150
0
1000
2000
3000
4000
5000
0
500
1000
0
25
50
75
100
0
100
200
300
0
50
100
150
0
1
2
3
4
5
0
5
10
15
20
2014
2016
2018
2020
2022

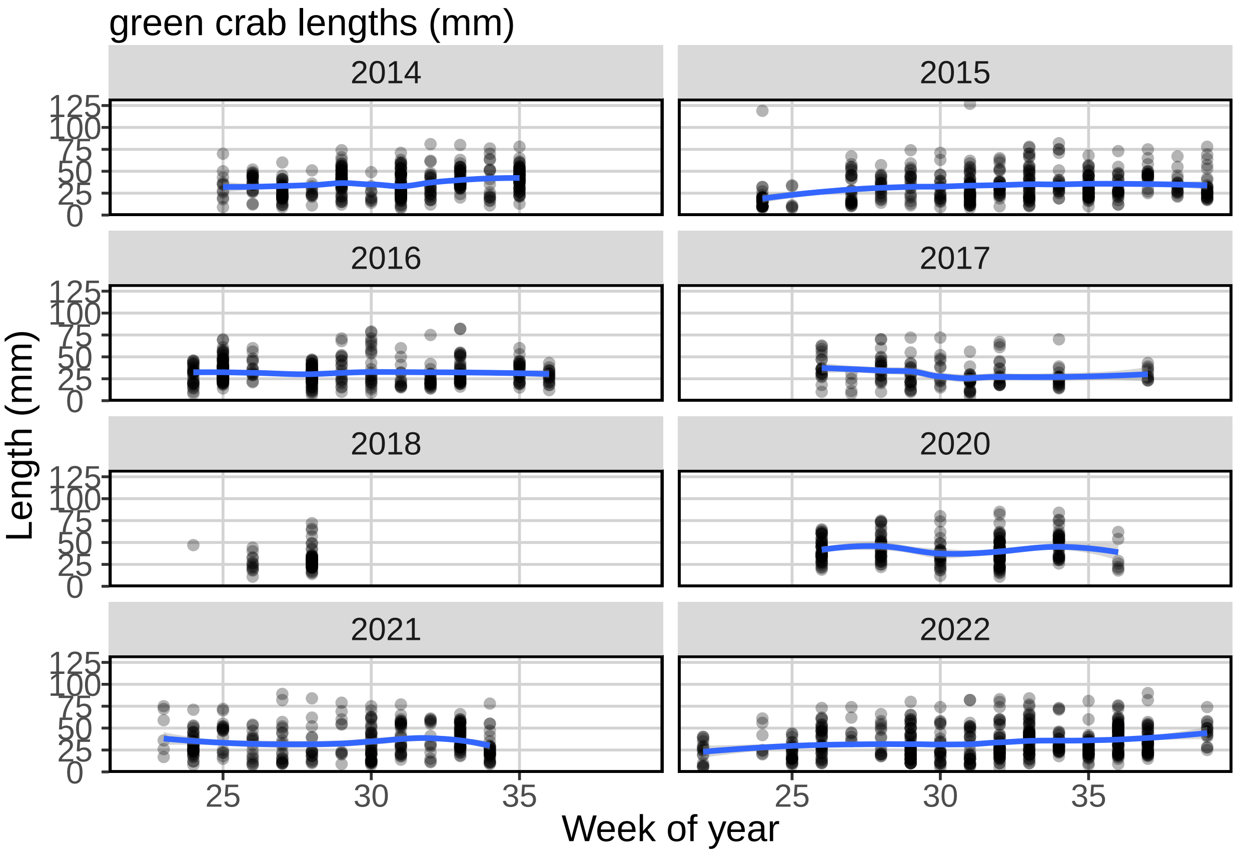
Size and growth

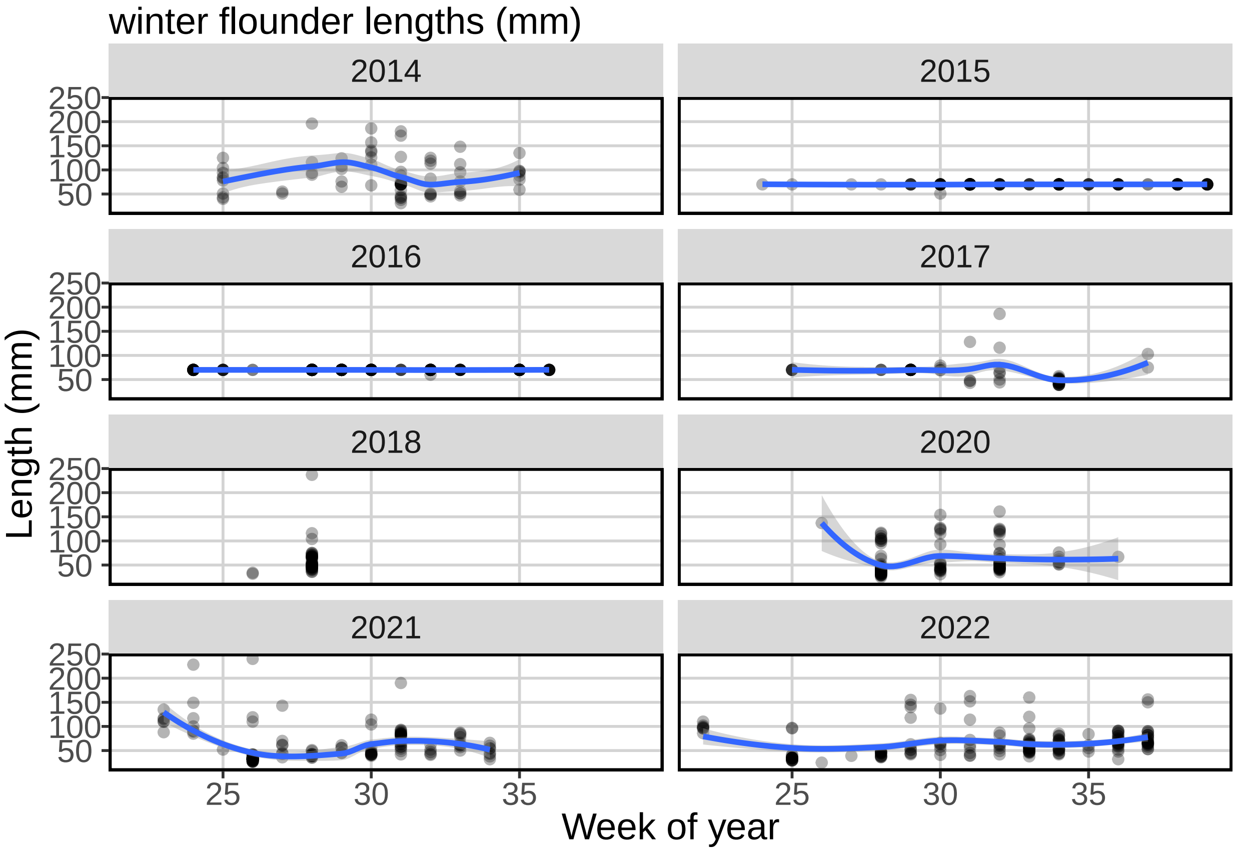
Next, we'll see if we can track growth of our most-encountered species by plotting lengths of sub-sampled individuals by week of the year.

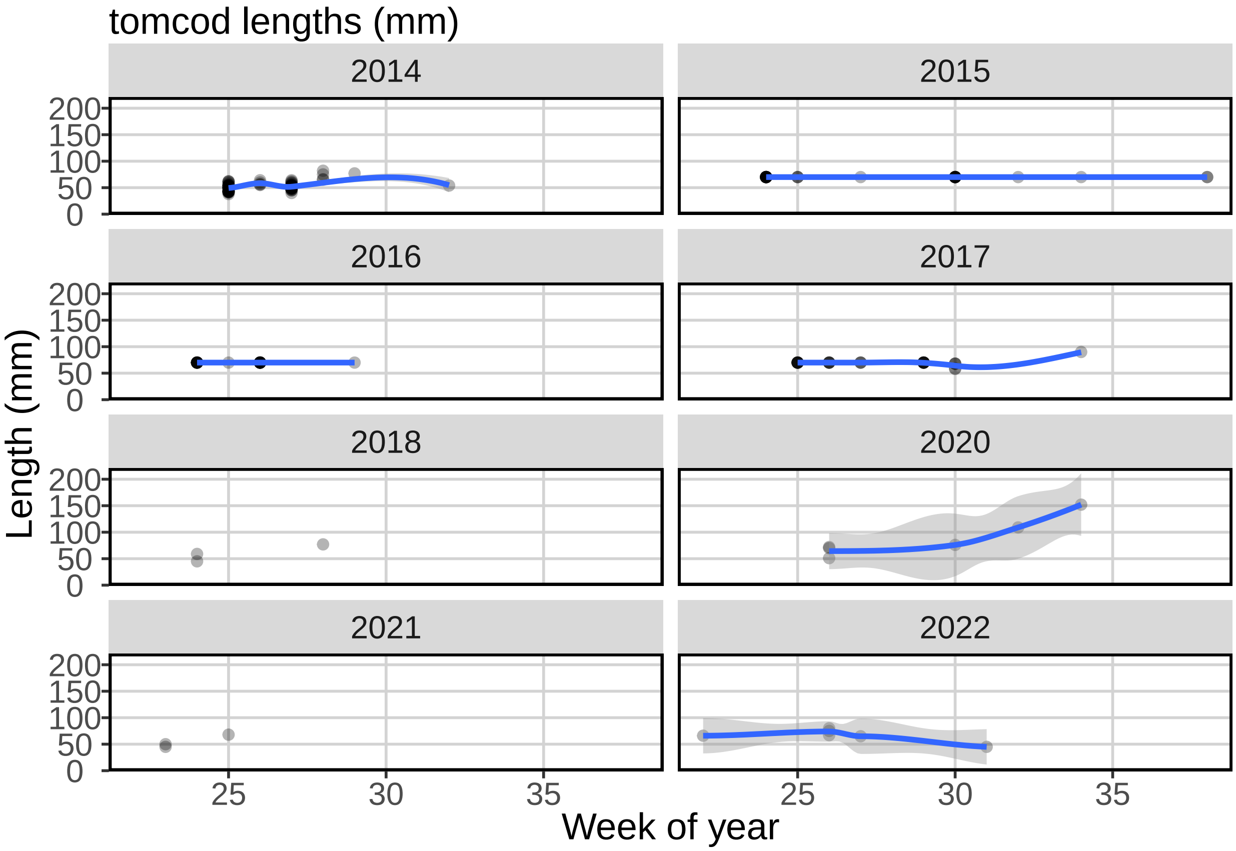
Honestly, the resulting plots look pretty terrible. It is likely that species with no clear growth over the summer period are constantly recruiting new individuals to the gear through growth of younger individuals or constantly moving new individuals into the nearshore area. This looks to be the case for green crabs and Atlantic silversides in particular.

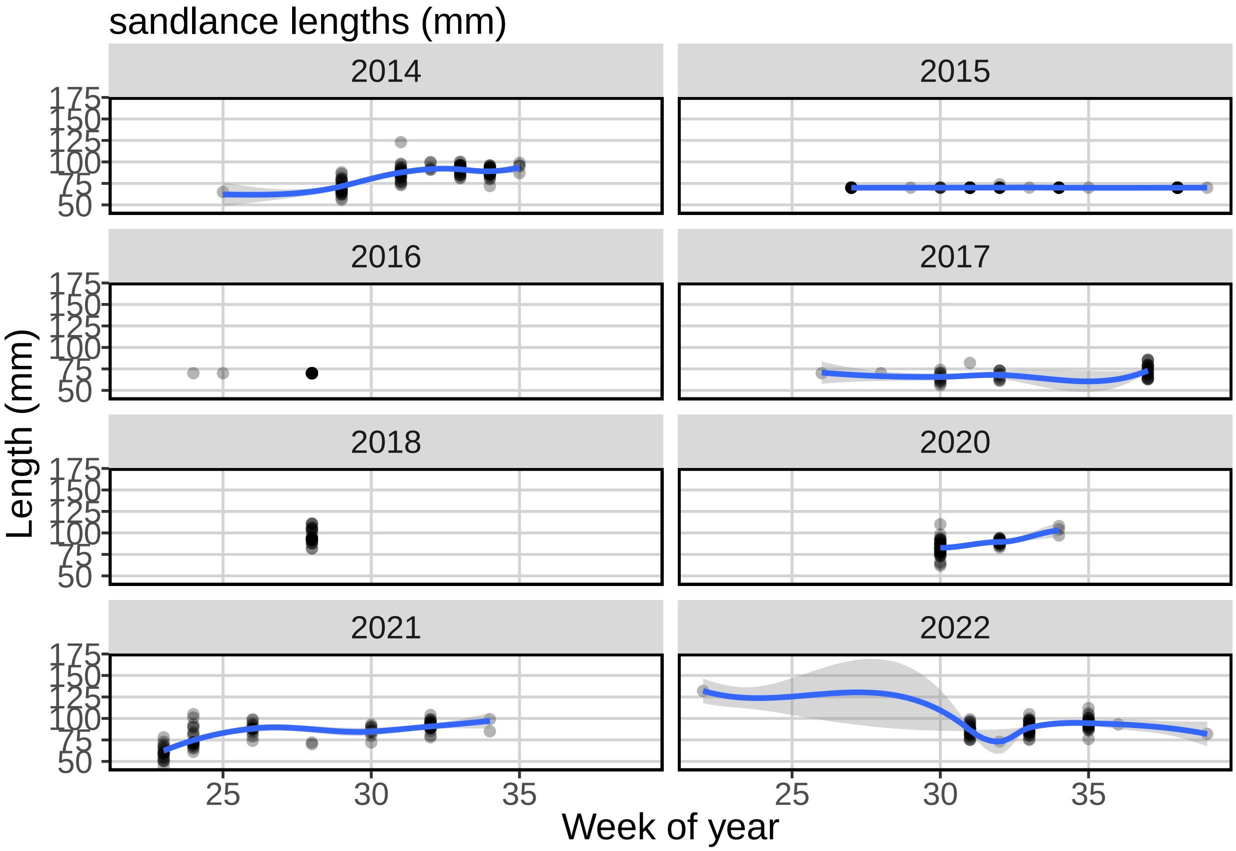
It's also possible that species with rapid declines in average length from early summer to mid summer are mixed cohorts– older individuals have overwintered in the nearshore area, are immediately available to the gear in June, and then it appears as though average size drops dramatically when the young-of-year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

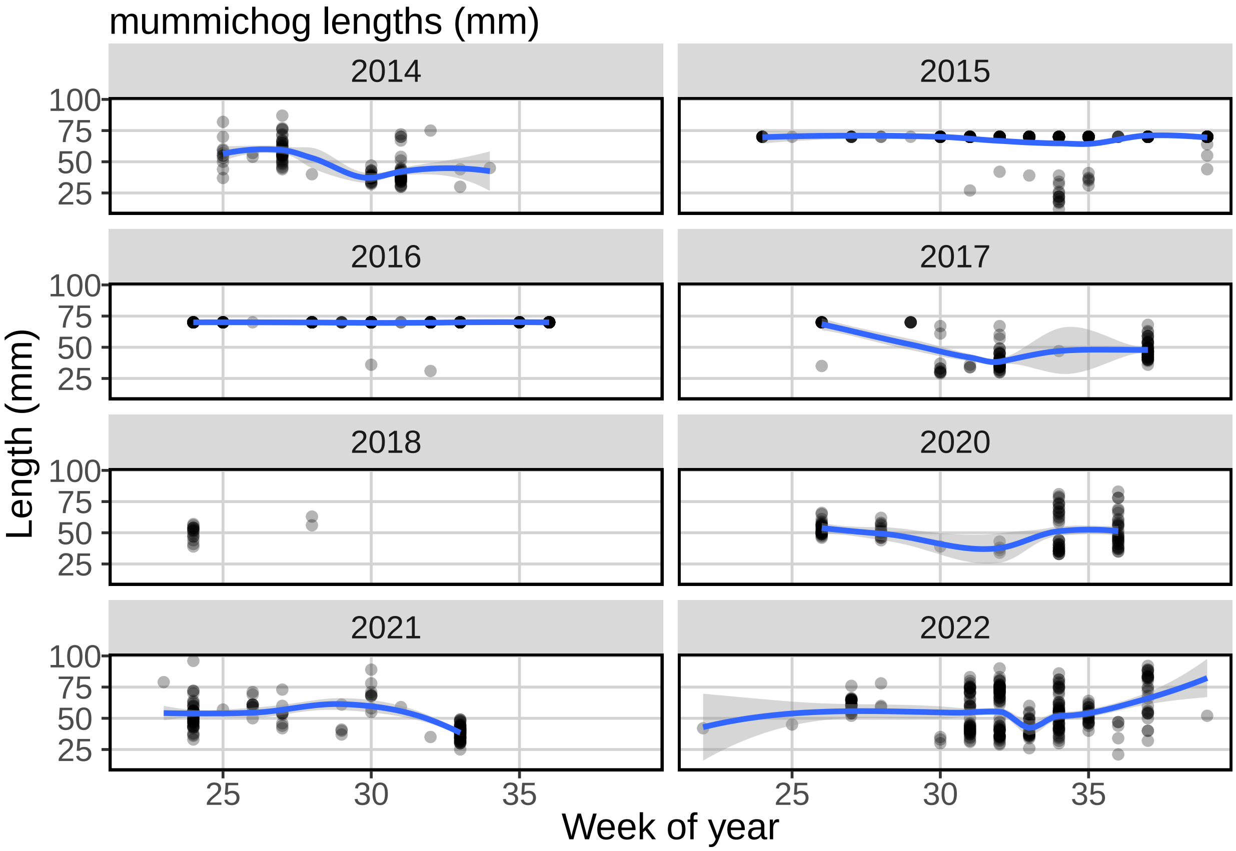
It would be beneficial to separate cohorts during analysis, if we want to look at growth rates. As is, we are still exploring and we'll keep everything together for now.

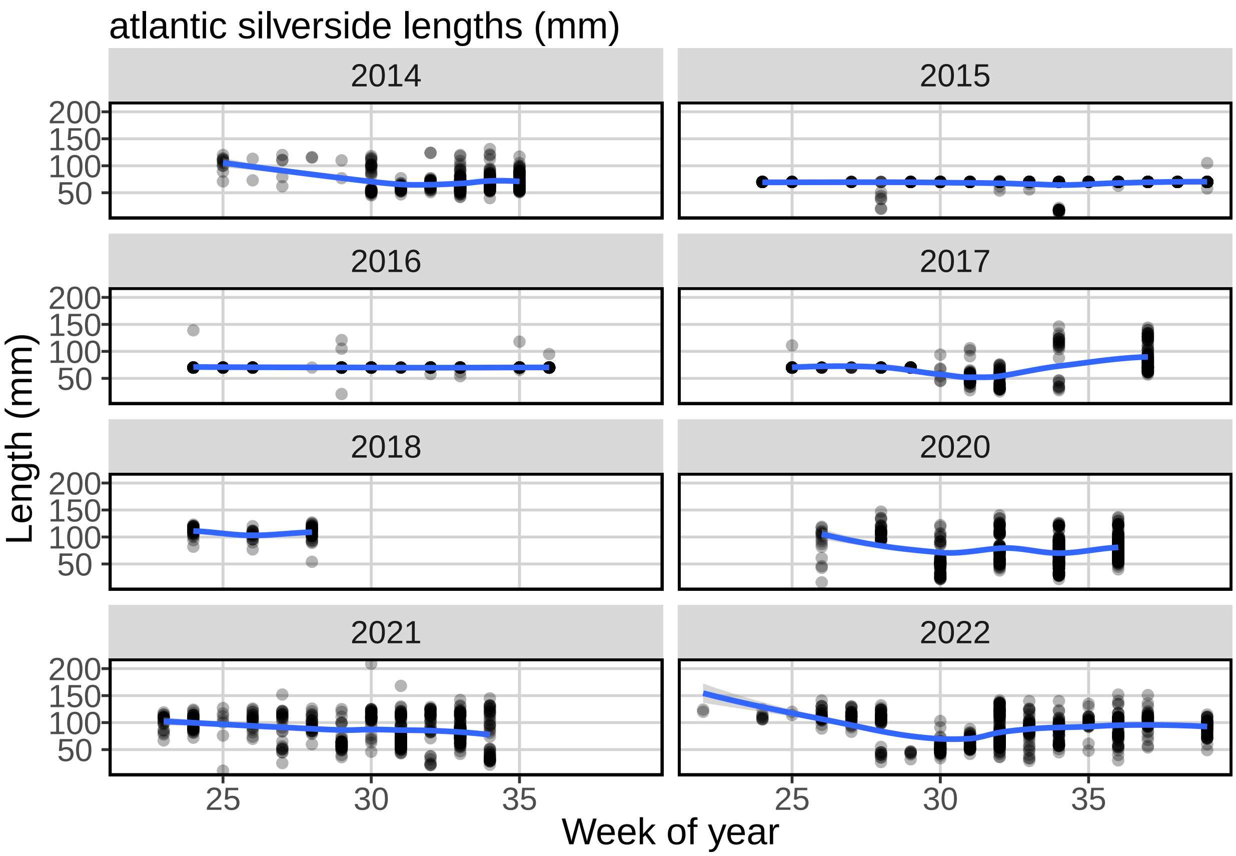


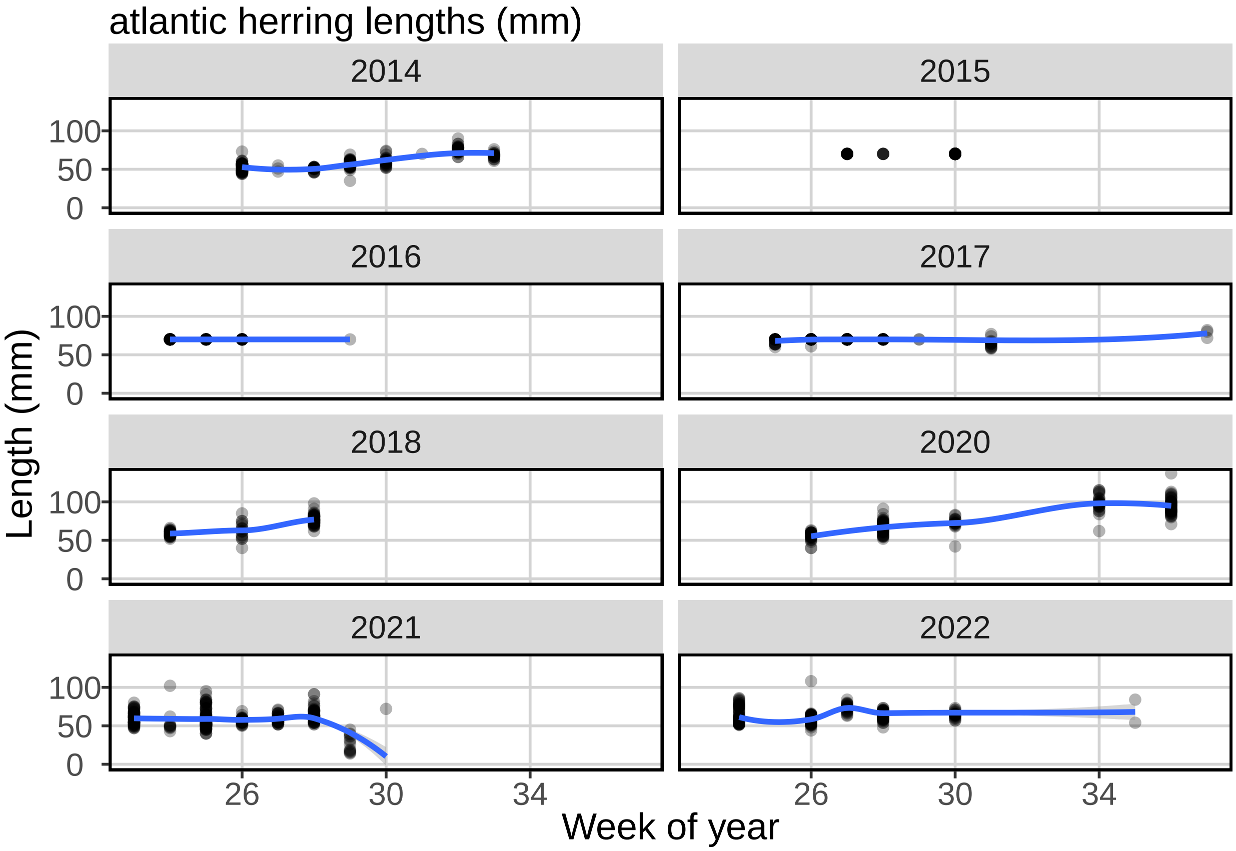

winter flounder lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
250
200
150
100
50
25
30
35
Week of year

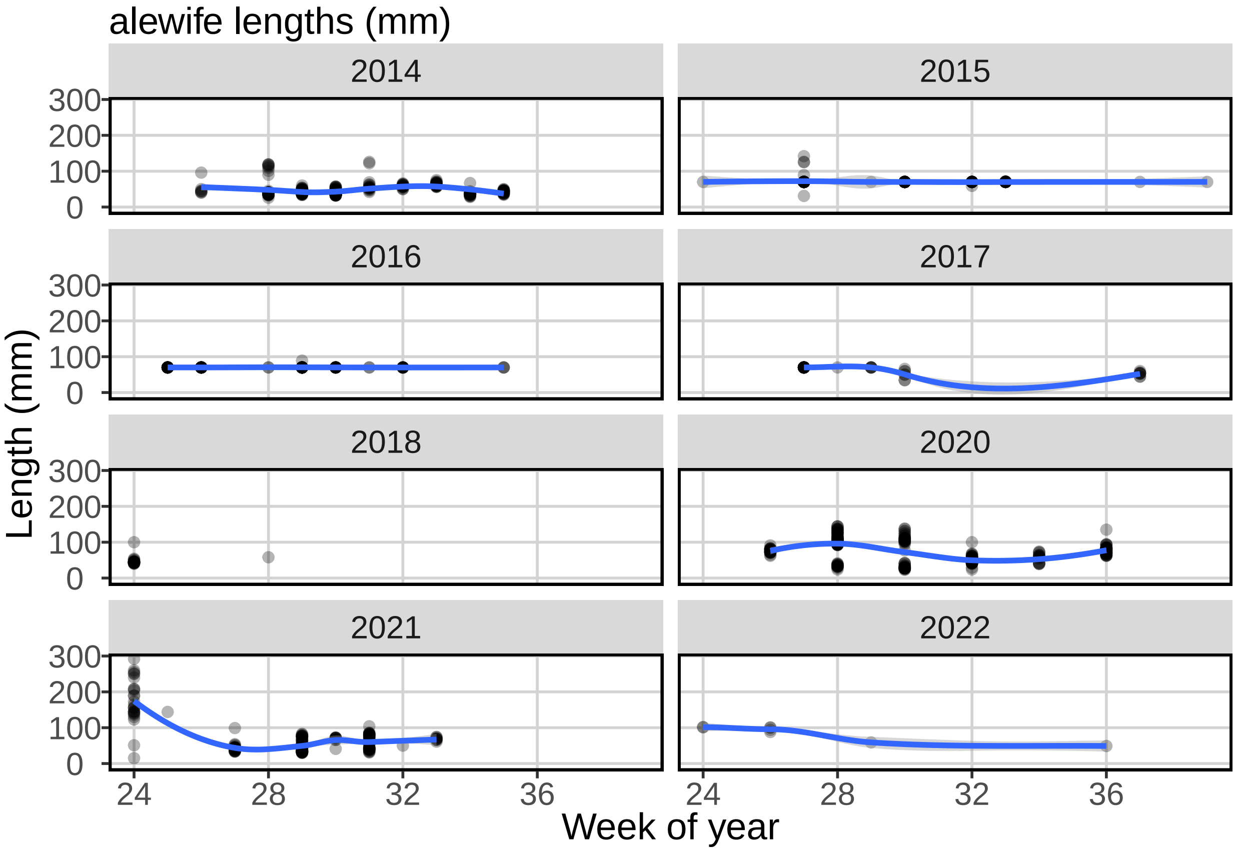

tomcod lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
200
150
100
50
0
25
30
35
Week of year


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
Week of year
25
30
35


atlantic silverside lengths (mm)
2014
2015
2016
2017
2018
2020
2021
2022
Length (mm)
200
150
100
50
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

Species that have similar environment requirements/ preferences should co-occur with each other frequently, and therefore may have some interesting interactions. Because we have raw abundance of each of our top 8 species and environmental covariate data (temperature, salinity, dissolved oxygen concentration) in every net haul, we can run a simple correlation analysis to check for any obvious links between species or species and their physical environments.

The results should be taken with a grain of salt. A simple correlation analysis like this does not take spatial, temporal, or spatio-temporal autocorrelation between observations into account.

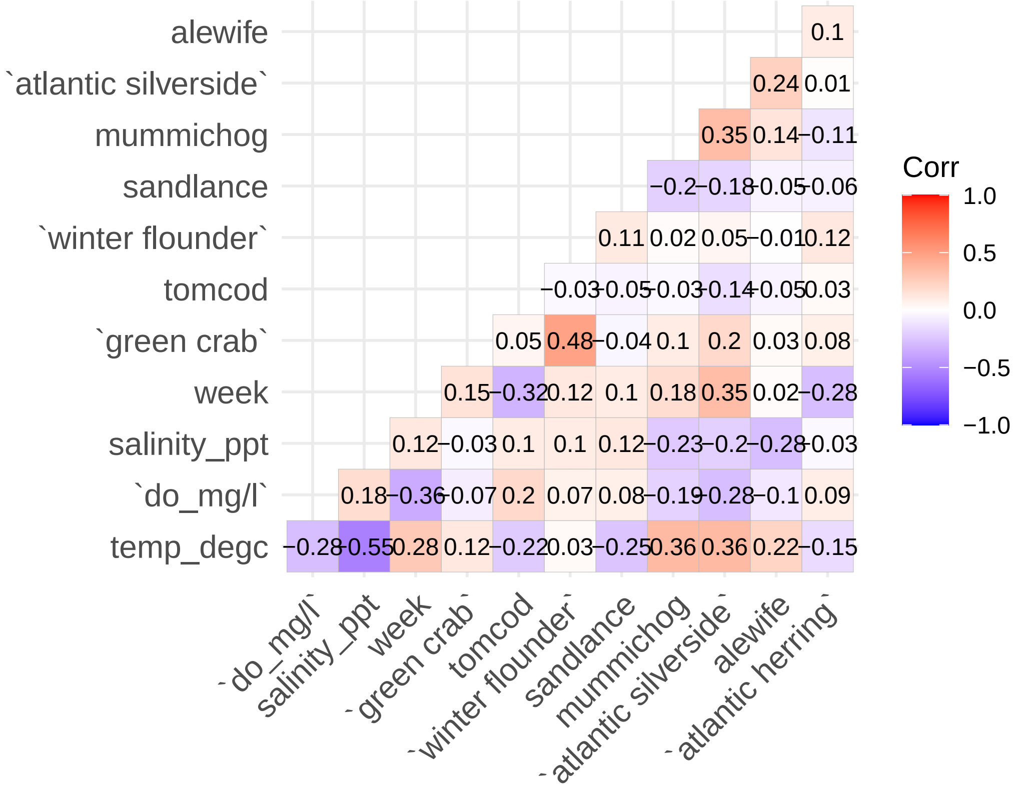


Figure 10: Density covariate correlation matrix

The rule of thumb is to consider relationships with an absolute correlation value of greater than 0.3 weak, greater than 0.5 moderate, and greater than 0.7 strong. We do not have any strong correlations. The only moderate correlation is between temperature and salinity, where increasing temperature is often correlated to decreasing salinity. This is expected in the context of our study area, which includes colder saltier water in the outer bay areas and warmer fresher waters in the inner bay areas. We also notice a weak negative correlation between dissolved oxygen concentration and week of year, which indicates decreased dissolved oxygen in the later sampling period. This should also be expected.

Mummichogs and silversides have weak positive correlations to temperature. No other species have any relationship to the three environmental variables. Tomcod have a weak negative correlation to week of the year, indicating higher abundance earlier in the sampling period. Silversides have a weak positive correlation to week of the year, indicating the opposite relationship.

Two pairs of species have weak positive correlations: green crabs and winter flounder, and mummichogs and silversides. This likely indicates overlaps in habitat preferences and residence time in the inshore region.

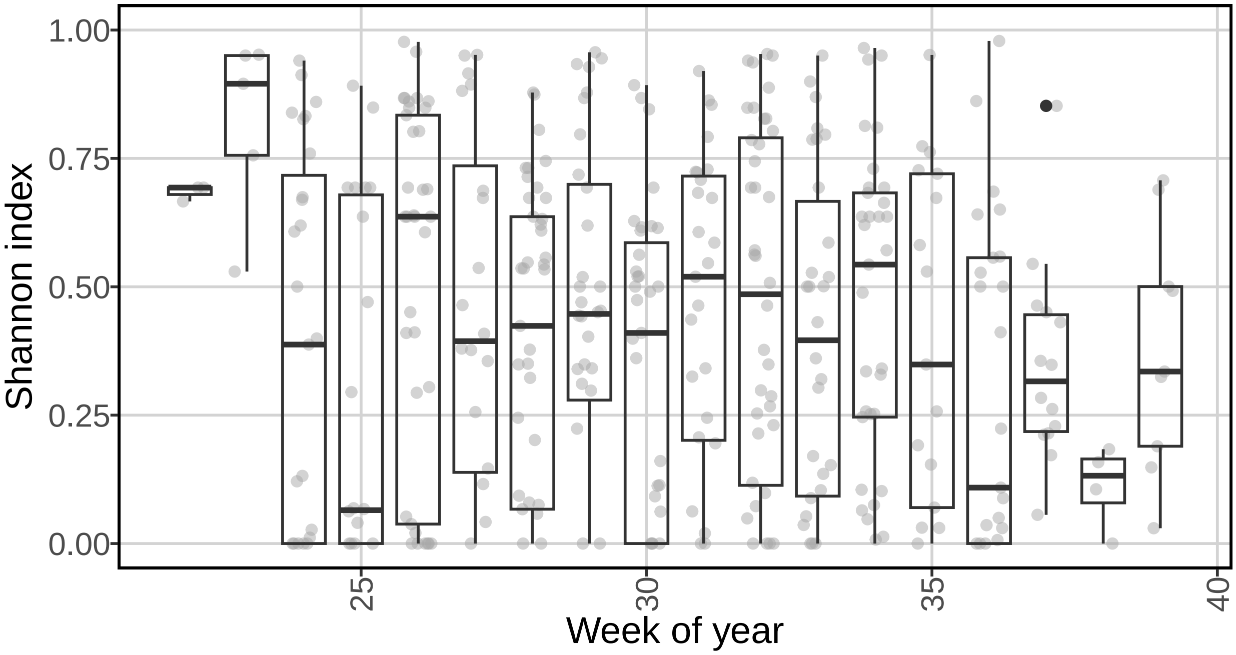
Diversity indices

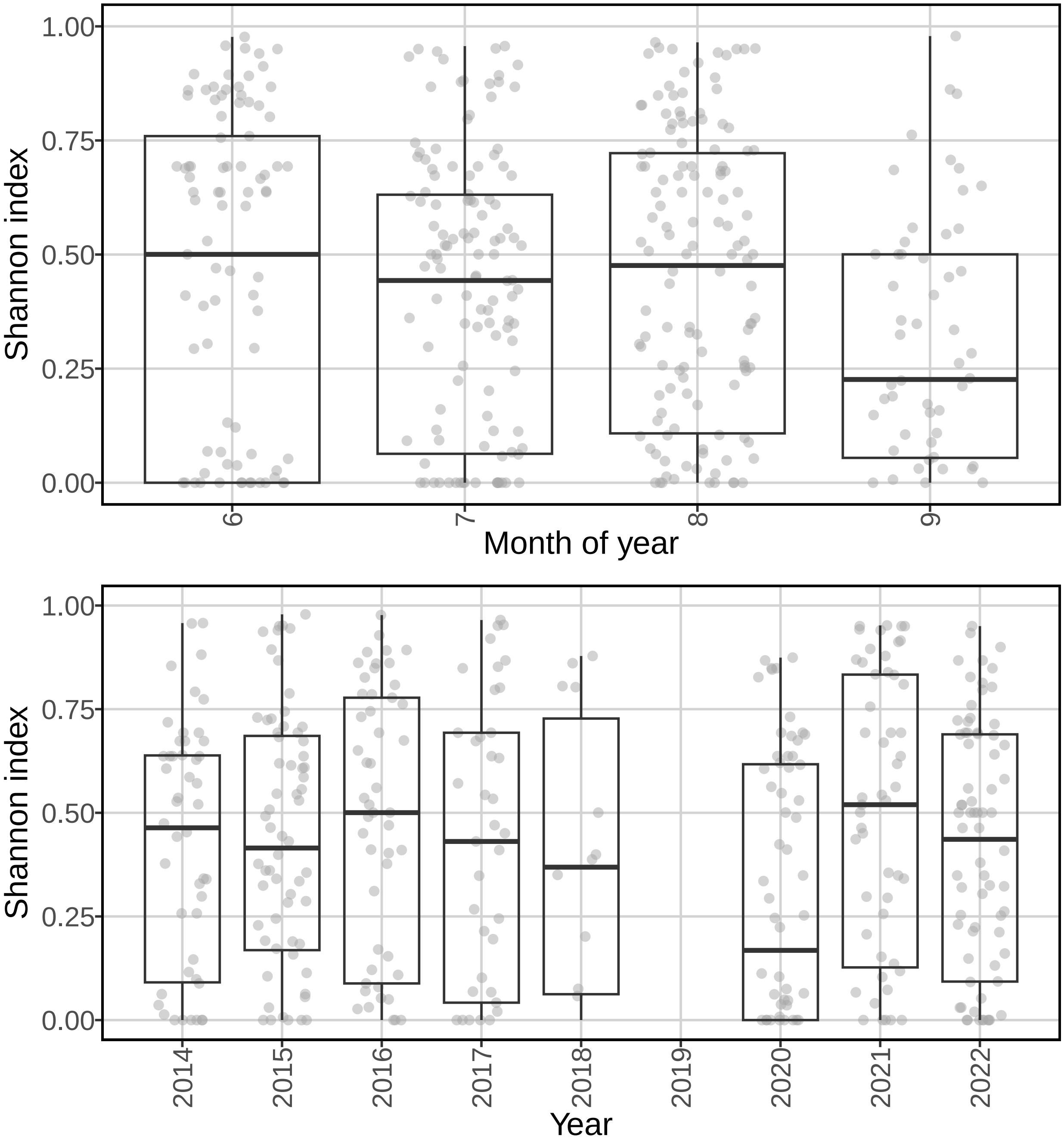
We can also examine spatial and temporal trends in common diversity indices, beginning with the Shannon index. The Shannon index is defined as

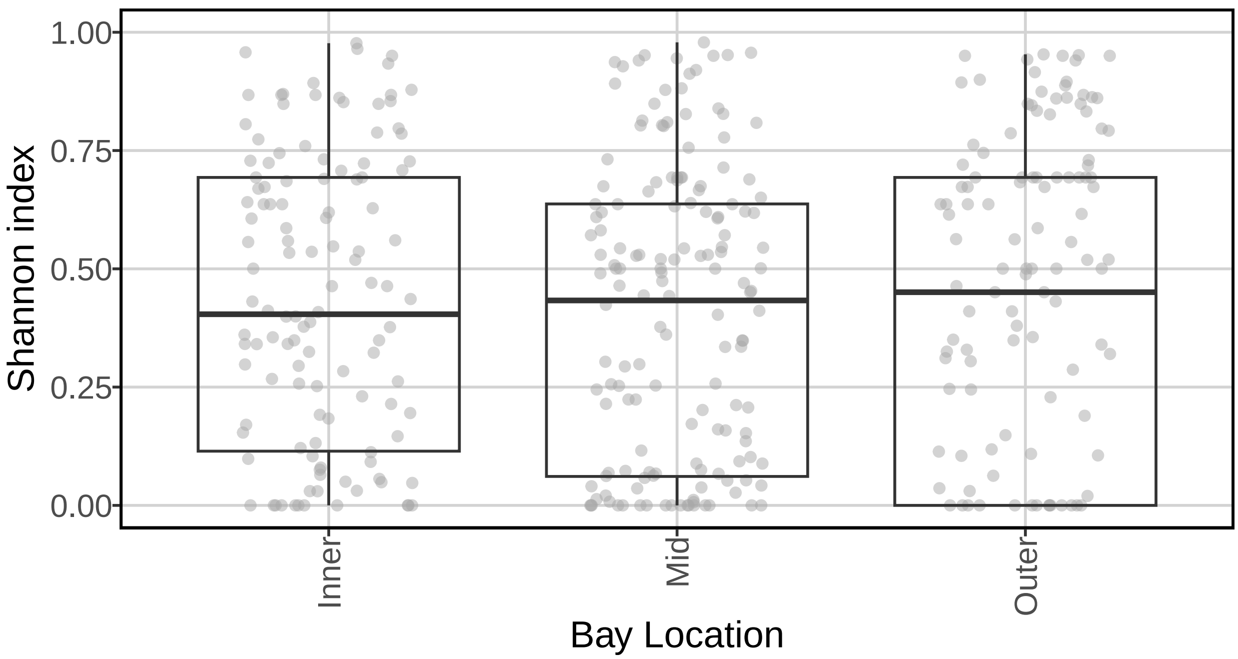
$$-\sum p_i * ln(p_i)$$

or the sum of the proportion of the entire community made up of species i multiplied by the natural log of the proportion of the entire community made up of species i. High Shannon index values indicate high biodiversity. An index value of 0 indicates that a community only has one species. The index is sensitive to rare species, and is generally better at providing quantitative measures of species richness. The Shannon index assumes random sampling and equal capture probability for all species in each net haul, which is likely a poor assumption for shore-based seines that occur in different habitats of Casco Bay.

We'll plot the Shannon index values over the weeks, months, years, and bay locations of the survey to identify any obvious connections.




Shannon index
1.00
0.75
0.50
0.25
0.00
6
7
8
9
Month of year
Shannon index
1.00
0.75
0.50
0.25
0.00
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year



It's difficult to find any clear trend in Shannon diversity index values within any of our possible groupings. I think if there were patterns to find in diversity, they wouldn't show up very well with the Shannon index. It's too sensitive to uncommon or rare species.

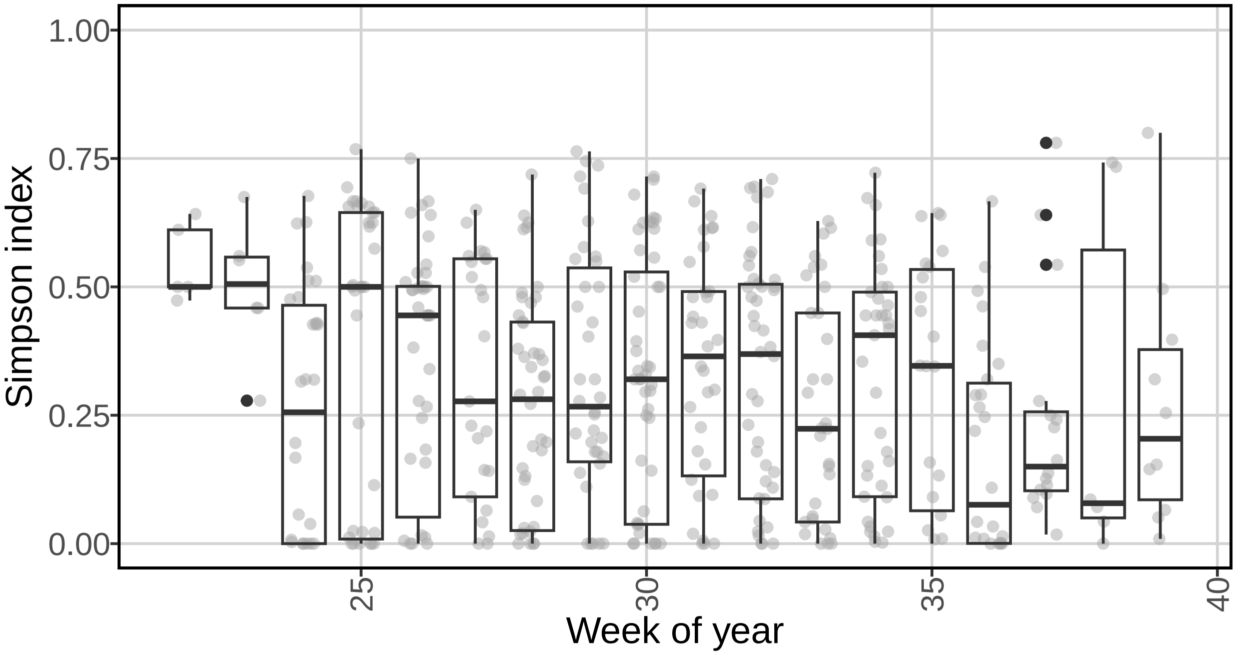
There are some notable outliers; the index value is abnormally high for week 23 and abnormally low for week 39. This is likely driven by very few samples occurring during those weeks. Median Shannon diversity index for 2020 is lower than all other years, but this may also be due to a pandemic-altered sampling schedule (all 12 sites in a week, as opposed to 6 sites in each week). There are no clear differences in diversity among the different bay locations or months.

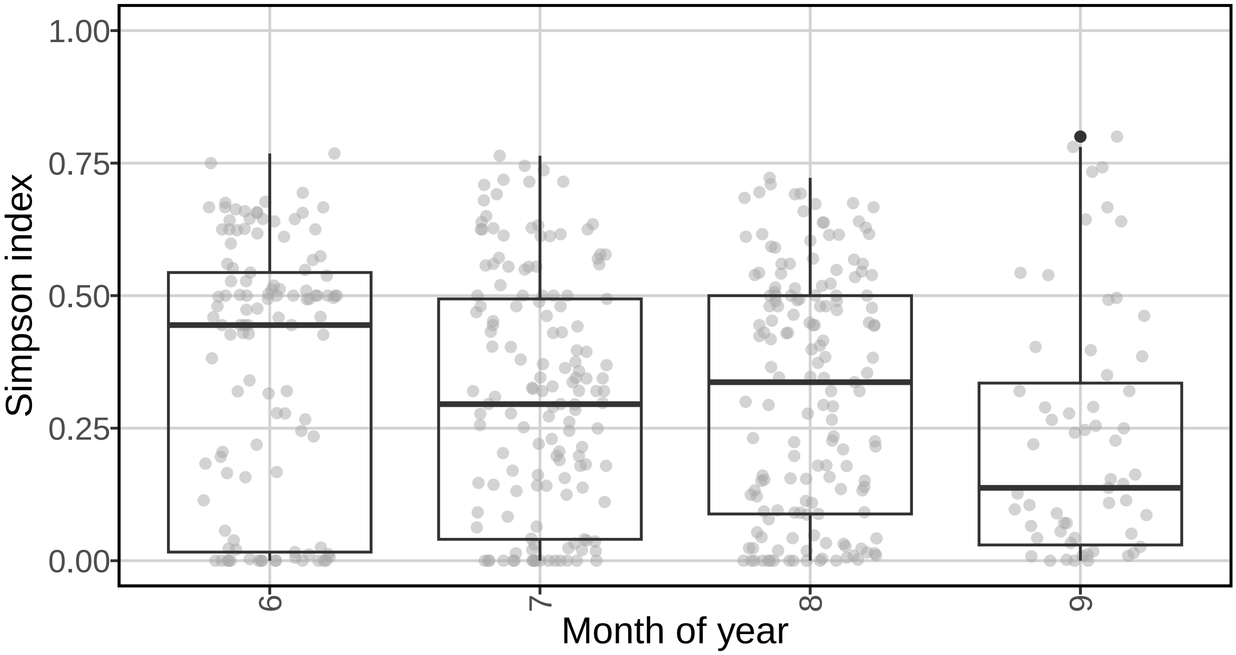
The second diversity index is the Simpson index. This index is defined as

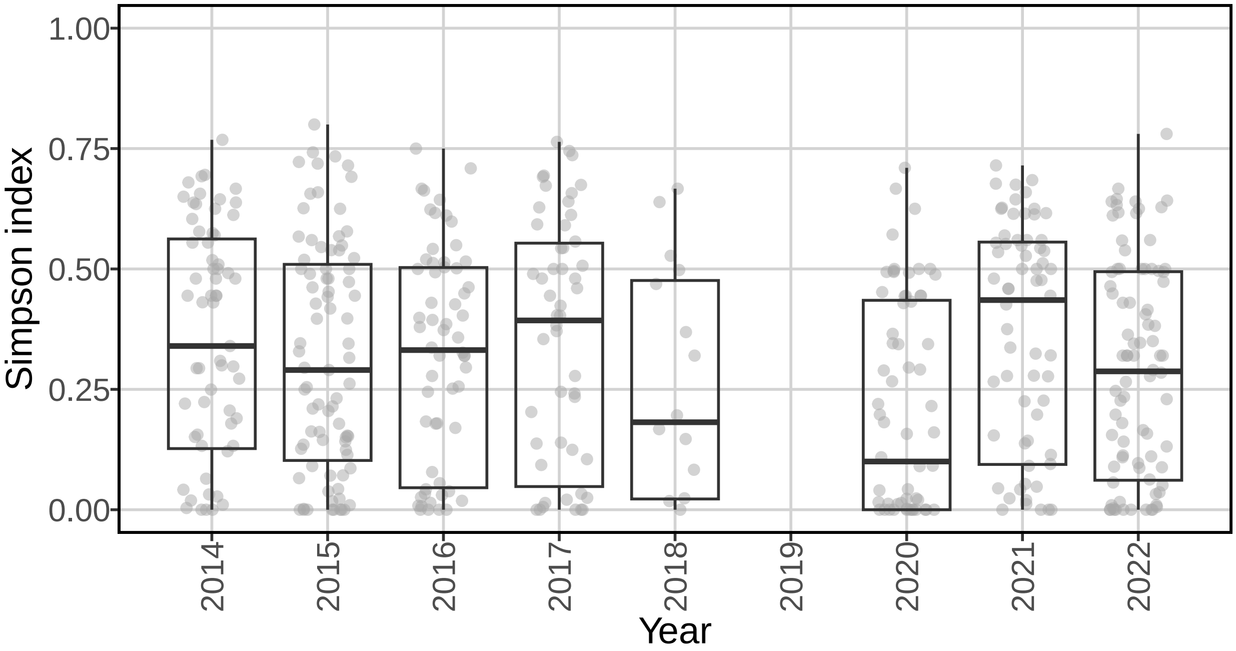
$$1 - \sum {p_i}^2$$

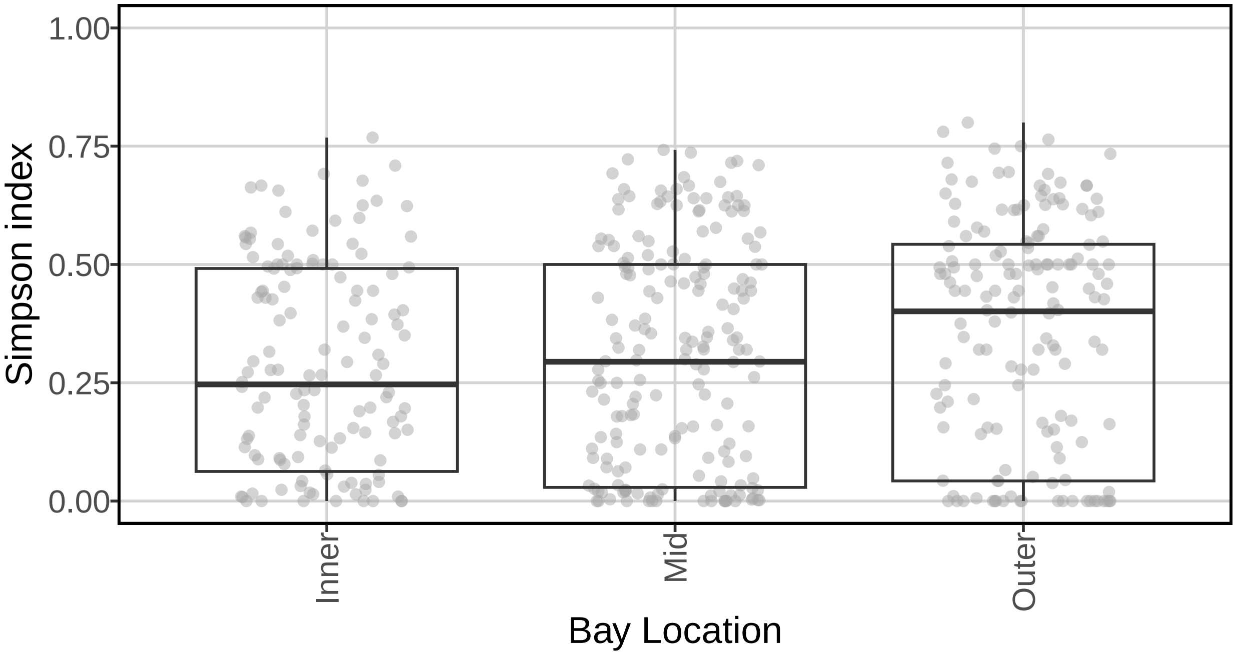
or 1 minus the sum of the proportion of the entire community made up of species i squared. The index will range between 0 and 1, with 0 indicating only one species was caught and 1 indicating high diversity. The Simpson index is not as sensitive to rare species, and does a better job of quantifying relative abundance.

We'll also plot the Simpson index values over the weeks, months, years, and bay locations of the survey.









These plots do not indicate much of a seasonal trend in diversity, either. There may be a slight decline in diversity though the sampled weeks, but this is not reflected in the plot of diversity among months. There is notably lower diversity in later summer than early summer, though. Thinking back to our plots of abundance, this makes sense. There are several species that leave the study area mid-summer and do not return, while there are fewer species that arrive during and remain in the study area for the late summer.

Like the Shannon index plots, the Simpson plots also highlight 2020 as a year with particularly low diversity and do not indicate any difference in diversity among the bay locations.

Conclusions

There are some cool stories to highlight here. Just some food for thought:

- Green crabs and juvenile winter flounder have ubiquitous and large populations within Casco Bay. They have similar habitat preferences and a weak correlation to each other. Their relative abundances have not changed much over time (which, in the case of winter flounder, is probably exciting for fishers).

- If you only look at abundance indices, tomcod appear to have a declining population. This is in line with various other organizations' published information on tomcod population health. However, the residence time rasters hint that there is more to the story; they may have also shifted seasonal use of the extreme nearshore area to earlier in the summer/late spring. Tomcod are short-lived, anadromous, benthopelagic fish. They typically spend their entire lives in estuarine and neashore areas. Though as shallow-water residents they are resistant to temperature swings, they prefer colder waters. Previous studies on temperature-linked distributions of juvenile tomcod have indicated both that they used to be abundant in nearshore Maine waters into July and that they are likely to avoid areas with bottom temperatures that exceed 22°C (Targett & McCleave 1974). Though we did not identify a correlation between tomcod abundance and surface water temperature, there is a significant difference between mean surface temperature when tomcod were present and mean surface temperature of all seine hauls. Mean water temperature for instances of tomcod capture was 1.6 degrees cooler than the mean temperature of all samples (p=0.0002, I did this test outside this document). I think it would be cool to begin surveying earlier to identify if we can identify a potential shift. I don't think we need to take bottom temperatures, as we're only seining in about 6 feet of water and there should be a tight coupling between surface and bottom temp.

- Similarly, temperature seems to be important for herring presence. Though again, no correlation was found between temperature and abundance, t-tests indicate significant differences between mean water temperature when herring are present and mean water temperature of all seine hauls (p=0.008). Mean surface temperature when herring were captured was 0.7 degrees cooler than mean surface temperature of all seines.

- Sandlance are important forage for many Gulf of Maine species. Their population dynamics within Casco Bay are likely impacted by myriad factors at much wider spatial scales, as reflected by the highly variable nature of the CBASS-generated indices of abundance.

- Atlantic silversides like heat. Their abundance has a weak positive correlation to temperature, and their abundance peaked in late summer during the year with the hottest mean temperature (2020). There is a significant difference between mean surface water temperature when silversides are present and mean surface water temperature of all seine hauls; silversides were present at waters on average 0.85 degrees hotter than our mean temperature. We can expect them to do well as water temperatures continue to warm, which is good because they can serve as forage for many other piscivorous species in the region.