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`. This sheet naming convention is kept for all Excel workbooks containing seine data. The data for 2022 are stored in the Excel workbook `2022_Raw_Seine_Data.xlsx`. The data for the latter half of 2018 and a few trips in 2019 are stored in `Missing_2018_2019_Seine_Data.xlsx` and were entered in 2023. There are only 2 trips noted for 2019, which may be fewer than actually took place. The data for 2023 are currently being entered and will be included in a later update.

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 QBC data You may or may not want to include data collected from the Harpswell side of Casco Bay. Toggle those inclusions in the next chunk of code.

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.

Load and clean 2018-19 data Some data from 2018 and 2019 was not entered in the digital spreadsheets until 2023. These observations will be added to the final dataframe.

Katie- edit

The species and sites dataframes do not need to be cleaned. The remainder need a quick look.

The trip information will be cleaned similar to the 2014-2021 dataset. Freshwater sites do not exist in these years. 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.

Load and clean 2023 2023 data were digitally entered in November and can now be included.

The species and sites dataframes do not need to be cleaned. The remainder need a quick look.

The trip information will be cleaned similar to the 2014-2021 dataset. Freshwater sites do not exist in these years. The site numbers and loc_id variables will be merged to match the patterns that exist in the 2014-2021 dataset.

```
##         mud  mud/gravel    mud/sand   mud/shell sand/gravel  sand/shell
##          13          13          14           6          25           7
```

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 all datasets to make a full record of this project.

Save data

Combined data can be saved as a new master seine spreadsheet.