Data Extraction Notes

General overview

1. The code imports data from the EMA program at Auke Bay lab. The EMA data are from the northern Bering Sea surveys and BASIS surveys not housed in the EcoDAAT database. These data were not sorted in Poland and were sorted under a variety of different persons and protocols. I spend time in the code converting these rather messy data to match the EcoDAAT structure. There is double counting of taxa of the same stage from the same station in these data, which I account for in the code below.
2. EcoDAAT data have been collected by EcoFOCI since the 1990s, but as you know it has significant spatial and temporal variability. Keep this in mind. We also had protocol changes in how the samples were processed over time, these are represented by different SPECIMEN\_FORM codes in the data set.
   1. Up to 2011, we only identified specific taxa
   2. In 2012, Jeff Napp changed the protocol to identify all taxa in the samples, but made some decisions about which taxa come from which nets that were different from what came prior, so particular taxa must be selected from particular nets to build a consistent data set over time
   3. I introduced a new protocol in 2019 to correct the data moving forward.
   4. Gear description found in GearDescription.xlsx

Overview of chunks in the extraction code

1. Load packages
2. Import EMA data
   1. Import EMA data (BASIS and north Bering Sea cruises not housed in EcoDAAT database. The raw data are in the Raw-Data folder.
   2. These data were collected by the EMA program at ABL and all methods questions should be directed to EMA staff.
   3. Combined data into one Raw file and export as .csv: EMA\_Combined\_RawData.csv
3. Convert EMA fields into EcoDAAT fields
   1. This code takes the EMA data and converts the information to match the format of the EcoDAAT data
   2. Export this file as .csv: EMA\_Combined\_RawData\_Recode.csv
4. Import EcoDAAT data
   1. This code connects to the EcoDAAT Oracle database and imports all available zooplankton data
   2. One needs a username and password to connect to the EcoDAAT database, so this will not work for those without access. Therefore, the code exports the full, raw data file so the subsequent code may be run: EcoDAAT\_RawData.csv
5. Do some data tidying for the EcoDAAT data set and combine the EcoDAAT and EMA datasets.
   1. Correct errors in the EcoDAAT mesh codes
   2. Check to see if duplicate cruises are in the EMA database and the EcoDAAT database and remove them.
   3. Combine two datasets prior to further processing and write combined raw file: EcoDAAT\_EMA\_Combined\_RawData.csv
6. Quick map of the raw data
   1. This quick map shows all of the data coverage prior to filtering
7. Begin filtering to match temporal, spatial ranges requested.
   1. Filter for all data north of 60°N and west of 155°W
   2. Remove 1996 from the dataset (standalone year)
   3. Gear differences
      1. 60BON: all years, two mesh sizes: 333 and 505 um
      2. 20BON: 2005-2021, mesh size 153 um
      3. Juday: from NBS survey from 2003-2007, 2009-2011, vertically towed net, 168 um mesh
      4. LG-CB: Arctic from 2010-2016, used with sled as oblique tow of smaller net,
      5. PairoVET: vertical tow, 153 um mesh, used in 2006, 2007, 2009, 2010 in NBS
8. Create a taxa list for annotation. This list will be used to lump taxa and determine which sizes and stages are selected from which nets. This annotated file, NBS\_Zoop\_TaxaSummary.csv, will be used to build the processed data set.
   1. First propose lumping categories as NBS\_Zoop\_CollatedTaxa.csv
9. Build *Calanus* dataset
   1. Filter for *Calanus* species
   2. Select the correct gear and mesh combinations for each dataset based on changing protocols and data source
   3. Add in zeros for positive catch stations
   4. Add biomass information
10. Build *Neocalanus* *cristatus* dataset
    1. Filter for *Calanus* species
    2. Select the correct gear and mesh combinations for each dataset based on changing protocols and data source
    3. Add in zeros for positive catch stations
    4. Add biomass information
       1. Note that adult *Neocalanus* *cristatus* are rare and I had not direct measures from Hopcroft, so I estimated using prosome length attributes (mean of all prosome lengths reported at WORMS) and converted to wet weight using a regression equation from (Kobari et al. 2003). You might wish to delete this since it is not a direct measure, but I included so you wouldn’t have gaps in the biomass estimates
11. Build *Neocalanus* spp. data set
    1. Filter for *Calanus* species
    2. Select the correct gear and mesh combinations for each dataset based on changing protocols and data source
    3. Add in zeros for positive catch stations
    4. Add biomass information
12. Write final data set
    1. Species\_Processed\_Final.csv, in Processed Data folder.