Laura Twardochleb

1/10/20

Data-paper read-me for R scripts and .csv files

**Occurrence data cleaning scripts on GitHub**

Scripts are available at <https://github.com/aquaXterra/aquatic_insects/tree/master/code/Data_cleaning_WQP_traits>

* “WQP\_occurrence\_cleaning.R” conducts initial data cleaning on biological monitoring data downloaded from the Water Quality Portal and assigns names and ranks to taxa. It requires the following files:
* macroinvertebrate\_biological.csv
* macroinvertebrate\_site.csv
* invertebrate\_biological.csv
* invertebrate\_site.csv

This script writes also writes two intermediate files: “WQP\_invertebrate\_1979\_2016\_2.csv” and “WQP\_insect\_w\_ranks.csv”. Final output is “insects2001v3.csv”.

* “State\_cleaning.R” conducts cleaning of raw state data files. It reads all .csv files in the folder “state\_data”. The script outputs “state\_data.csv”.
* “WQP\_state\_merge.R” conducts remaining cleaning of occurrence data. The script requires the following files:
* state\_data.csv
* insects2001v3.csv
* raw\_trait\_table.csv
* The script writes “occurrence22.1.csv”, which is the final cleaned occurrence dataset with genus only records for Ch. 3 analysis paper. It also writes “ancillary\_taxonomy\_occurrence.csv” to create the ancillary taxonomy table. The script writes “occurrence22.csv”, which contains all of cleaned occurrence data, and “Genus\_Occurrences.csv”, which contains the genus occurrence table for the data paper and hosting by EDI. Finally, it writes, “Raw\_Community\_Data.csv”, which is the raw occurrence table for EDI.

**Trait data cleaning scripts on GitHub**

* “WQP\_insect\_2001.R” conducts initial cleaning of USEPA traits database and outputs two files, “biotraits\_11\_30\_17.csv”, which contains cleaned data from the USEPA traits database, and “biotraits8.csv”, which contains taxa not in USEPA database for which Ethan Hiltner added trait data. The script requires the following files:
  + - * insects2001v3.csv
      * freshwater\_biotraits\_transposed.csv
* “Pyne\_traits.R” combines and harmonizes data between Matthew Pyne’s trait dataset and the USEPA traits database. It reads the Pyne data and the USEPA database that was added to by Ethan Hiltner:
* biotraits8\_6\_15\_18.csv
* Pyne2016\_genus.csv

The script writes the data marge to “biotraits\_8\_23\_18.csv”. It also writes to subsets of the trait dataset to file for Ethan Hiltner, Minali Bhatt and Erika Ralston to work from: “Hiltner\_biotraits.csv”, “Ralston\_biotraits.csv” and “Bhatt\_biotraits.csv”. Further modifications were made in the script to Bhatt and Ralston subsets of the data and output to “Ralstin\_biotraits\_11\_16\_18.csv” and “Bhatt\_biotraits\_11\_16\_18.csv”.

* “Trait\_cleaning\_12\_10\_19.R” conducts remaining data cleaning of traits database by combining all subsets of the database from Ethan Hiltner, Minali Bhatt, and Erika Ralston. It requires Bhatt, Hiltner, and Ralston .csv files, with code to read them in from Google Drive. The script outputs “biotraits\_wide2.csv”, which contains the cleaned, raw trait data.
* “Calc\_trait\_affinity\_mode.R” calculates trait affinities and modal trait values for two tables: Genus\_Trait\_Affinities (“Genus\_Trait\_Affinities.csv”) and Genus\_Traits (“Genus\_Traits.csv”). It also produces an Ancillary\_Taxonomy\_Table (“Ancillary\_Taxonomy.csv”) and the Raw\_Traits table (“Raw\_Traits.csv”) formatted for EDI. It requires the following files:
* biotraits\_wide2.csv
* ancillary\_taxonomy\_occurrence.csv
* epa\_traits.csv
* epa\_citations.csv
* Raw\_Traits copy\_Pyne.csv, which is Matt Pyne’s notes on his traits (created manually in excel)
* biotraits\_revision4.csv, which is the trait table that I manually corrected for problems with taxonomic names through manually searching ITIS.gov

**Data on HPCC**

Data files are available at /mnt/ufs18/rs-022/aquaxterra/DATA/Insects

* csv\_files contains the raw state data files in .csv format. These are needed for the script, “State\_cleaning.R”.
* WQP\_trait\_data contains the other files read by all R scripts for intermediate data processing.
* final\_data/Files\_EDI/Files\_EDI contains the final output files for EDI, including: “Ancillary\_Sample\_Method.csv”, “Ancillary\_Trait.csv”, “Calc\_trait\_affinity\_mode.csv”,
* Data\_Sources.csv”, “Genus\_Occurrences.csv”, “Raw\_Community\_Data.csv”
* final\_data/Files\_EDI/Files\_EDI/trait\_revision\_for\_EDI contains the revised trait data files for EDI, including: “Ancillary\_Taxonomy.csv”, “Genus\_Trait\_Affinities.csv”, “Genus\_Traits.csv”, “Raw\_Traits.csv”