Pelagic Larval Duration Random Forest README

The home range model I am proposing (this could merit additional lit review to check for other relevant parameters):

*Home range ~ beta0 + beta1(r) + beta2(K) + beta3(categorical movement parameter) + beta4(IUCN red list status)  + beta5(body length) + beta6(trophic level) + beta7(pelagic or demersal)*

Model

PLD(Family

Home range and PLD may be negatively correlated. Any of the pelagics have low PLD while sessile/low mobility species tend to have long PLD (e.g. lobsters).

Data:

We need pelagic larval duration information for **811 unique fish and invertebrates species.**

Marshall et al. reported life history data for 806 invertebrate species. 254 of which have “planktonic time” data. **2 matches our modeled species list.**

Ramesh et al.’s life history data table contains 1798 rows but there are duplicates (i.e., data compiled from several studies per species?). There are 265 unique entries with larval duration data. **65 matches our modeled species list**.

![Chart, histogram

Description automatically generated]()

![Chart, histogram

Description automatically generated]()

![Chart

Description automatically generated]()

![A picture containing graphical user interface

Description automatically generated]()

![Chart, scatter chart

Description automatically generated]()

To do:

* Check the families of our species list.
* Check how many species per family.

* Home range data collected via lit review can be found in mpa-fish-flows/data/01\_raw/homerange/all\_homerange.csv; column = ‘hr’
  + <https://drive.google.com/file/d/1nQM44mEgcsfe47QaJK2EwU9lBvs9O919/view?usp=sharing>
* For species list and model parameters, data can be found in the following locations:
  + species list, r, K, categorical movement parameter: MegaData\_Ray.rdf (available: <https://drive.google.com/file/d/1vrrs4SC-pJfWwNbWJGPyXERC7uZUMkQZ/view>)
    - species = SciName
    - K = Kfin
    - r = r\_fin
    - categorical movement parameter = m\_fin
* IUCN red list status: included\_aquamaps\_spp\_list.csv (available: <https://drive.google.com/file/d/1IOiK8s2jCiJ1jKv0ksYlRQti1fstNz3r/view>)
  + IUCN red list status reported in ‘iucn\_category’
  + note this has a different larger associated species list and so will need to be merged via a left join (merging by SciName in the prior file and valid\_sci\_name
  + the script with the original species matching for this exercise is here (if useful): https://github.com/emlab-ucsb/ocean-conservation-priorities/blob/master/data\_prep/01\_get\_species\_list.Rmd
* body length, trophic level, pelagic/demersal: available via Fishbase (see Prior Analyses section below for more details)

Relevant Github repos:

* mpa-fish-flows: <https://github.com/emlab-ucsb/mpa-fish-flows>
  + All analyses should go here, noting that there are not yet any scripts associated with any homerange analyses
* ocean-conservation-priorities: <https://github.com/emlab-ucsb/ocean-conservation-priorities>
  + This houses the original scripts used in Sala et al. 2021 (<https://www.nature.com/articles/s41586-021-03371-z>), one of which is linked above; it is a public repository

Prior analyses (completed by Tim White, former PhD student at Stanford) can be found in mpa-fish-flows/other-project-materials/homerange/tim-hrmodel

* The README document details his original approach, which we augmented with additional data lit review searches and expanded data collection (e.g. KUD SD in addition to mean)
* Tim’s original model used trophic level and body size to predict home range (there are scripts detailing how he pulled this info from Fishbase using the rfishbase package and how missing values were interpolated.

General approach:

1. Create data frame with model parameters: merge relevant columns from MegaData\_Ray.RDS with included\_aquamaps\_spp\_list.csv with all\_homerange.csv (retaining full species list from MegaData\_Ray.RDS)
2. Get Fishbase parameters from the rfishbase package; use Tim’s method reported in the read me to impute missing values
3. Sample training data based on family group (or some higher taxonomic classification or trophic level if training data are sparse)
4. Run RF model