User Manual for the R package FishFreqTree

Haikun Xu

11/29/2021

Introduction

This R package helps users to easily explore and quantitavely compare fishery definitions based on a distributional regression tree algorithm that is applied to age/length frequency data. The details regarding distributional regression tree algorithm can be found in Lennert-Cody et al. 2010 and 2013.

Input data

The input age/length frequency data should be a data frame including four columns named exactly as "lat", "lon", "year", and "quarter" and various length frequency columns corresponding to selected age/length bins. This regression tree package works with age/length frequency data so please make sure age/length frequency values sum to 1 across age/length bins. An example of the input data can be found here.

Functions

Main functions

`run_regression_tree`: explore a user-specified fishery definition based on the regression tree algorithm

Users are required to specify the input data frame (`LF`), the first (`fcol`) and last (`lcol`) columns in the input data frame that have frequency data, the name of all age/length bins (`bins`) as a numeric vector, the number of splits (`Nsplit`; equals to the number of defined fisheries - 1), and a directory where results are saved (`save_dir`).

The function also provides some advanced options including manually building the regression tree (`manual = TRUE`) and specifying the minimal number of lat (`lat.min`), lon (`lon.min`), and year (`year.min`) allowed for a cell. Users can also turn on/off the year (`year`) and quarter (`quarter`) dimensions when building the regression tree.

This function provides a series of standardized outputs for users to understand the result:

- split.csv: all candidate splits are compared and sorted across existing cells based on the percentage of total variance explained. This table is used to specify the stepwise selection decision.
- improvement-split.csv: cell-specific improvement metric for all candidate splits; values are sorted for every cell (highest values are preferred for the selection). This table provides supplementary information only and is NOT used to specify the step-wise selection decision.

- Record.csv: summarize step-wise split information including split number, key, value, cell, and the percentage of total variance explained.
- split(annual maps).png: spatial distribution of cells across quarters
- split(quarterly maps).png: spatial distribution of cells by quarter
- split(latlon).png: cell-specific improvement profiles against lat and lon
- split(lf).png: comparison of cell-specific length frequency

The package provides a default fishery definition (**`run_regression_tree(..., manual = FALSE)`**) by selecting every split that corresponds to the highest percentage of variability explained (the first row of split.csv files). However, users can explore other definitions by using **`run_regression_tree(..., manual = TRUE, select = user_specified)`**. The user-specified splits are numbered according to the rank in split.csv files). Please change one split at a time because the step-wise regression tree is hierarchical.

`loop_regression_tree`: compare differring fishery definitions according to the percentage of variance explained

Users are highly recommended to compare various fishery definitions, even for the same number of splits, because the definition is flexible and may need to be adjusted for practical reasons. Moreover, the tree is hierarchical and unstable, so comparing a variety of combinations with the default combination is highly valuable. In fact, the default one may not explain the highest percentage of variance in the input data.

Users are required to specify the input data frame (`LF`), the first (`fcol`) and last (`lcol`) columns in the input data frame that have frequency data, the name of all age/length bins (`bins`) as a numeric vector, the number of splits (`Nsplit`; equals to the number of defined fisheries - 1), a directory where results are saved (`save_dir`), and the maximal number of candidate splits explored for each split (`max_select`).

The function also provides some advanced options including specifying the minimal number of lat (`lat.min`), lon (`lon.min`), and year (`year.min`) allowed for a cell. Users can also turn on/off the year (`year`) and quarter (`quarter`) dimensions when building the regression tree.

The function changes the selection of one split at a time, from the first choice to the `max_select`th choice (the first `max_select` rows of split.csv files), through a loop. Then the function summarizes all explored combinations in a table with their percentages of variance explained, which can be used for comparison.

For example, there are three splits (`Nsplit = 3`) and `max_select` is specified to be 3. The function provides to the screen the summary table from the loop function (also saved as loop.csv).

select1	select2	select3	Var_explained
1	2	1	0.164171
1	1	1	0.161714
1	3	1	0.149531
2	1	1	0.145291
3	1	1	0.130918

Supporting functions