User Manual for the R package FishFreqTree

Haikun Xu

11/29/2021

Introduction

This R package *FishFreqTree* helps users to easily explore and quantitatively 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 Lennert-Cody et al. (2013).

How to install

```
library(devtools)
install_github('HaikunXu/FishFreqTree',ref='main')
```

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. The last column of this table (Rank) is used to specify the step-wise 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(year).png: cell-specific improvement profiles against year
- 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 differing 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

Select1	Select2	Select3	Var_explained
3	1	1	0.130918

Supporting functions

`make.lf.map`: make lat-lon gridded maps for length frequency

`make_meanl.map`: make spatial maps for mean length

`lf.aggregate`: aggregate length count data by length only or also by year and quarter

`lf.demean`: remove the mean of length frequency data

Package Demonstration

Load the package and example data (https://github.com/HaikunXu/FishFreqTree/blob/main/manual/LF.RData):

```
# devtools::install_github('HaikunXu/RegressionTree',ref='main')
library(FishFreqTree)
require(tidyverse) # it is required for some supporting functions
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

```
load(file = "D:/OneDrive - IATTC/Git/FishFreqTree/manual/LF.RData")
head(LF)
```

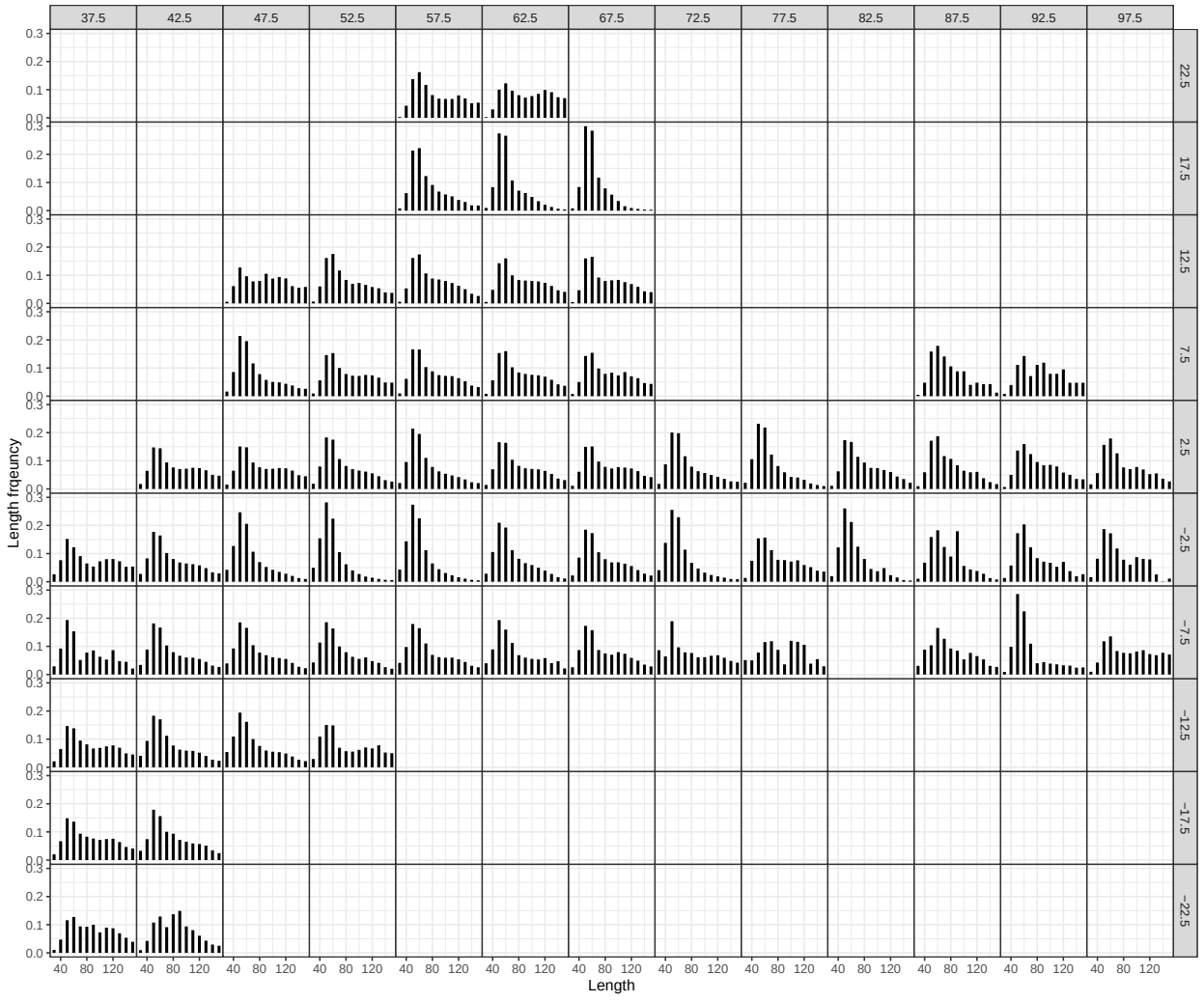
```
## # A tibble: 6 x 17
## # Groups:   year, quarter, lat, lon [6]
##    year quarter   lat   lon    `30`   `40`   `50`   `60`   `70`   `80`   `90`
##   <dbl>   <dbl> <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1    31       1  -2.5  52.5 0.0195  0.0637 0.108  0.140  0.138  0.121  0.0908
## 2    31       1  -2.5  57.5 0.0157  0.0705 0.125  0.154  0.133  0.0966 0.107
## 3    31       1  -2.5  62.5 0.0214  0.0646 0.0986 0.143  0.134  0.116  0.0919
## 4    31       1   2.5  62.5 0.0114  0.0496 0.101  0.136  0.145  0.126  0.0966
## 5    31       1   2.5  67.5 0.00759 0.0511 0.0926 0.141  0.155  0.124  0.0926
## 6    31       2  -2.5  52.5 0.0216  0.0594 0.112  0.0971 0.0989 0.115  0.126
## # ... with 6 more variables: 100Â <dbl>, 110Â <dbl>, 120Â <dbl>, 130Â <dbl>,
## #   140Â <dbl>, 150Â <dbl>
```

Specify function inputs

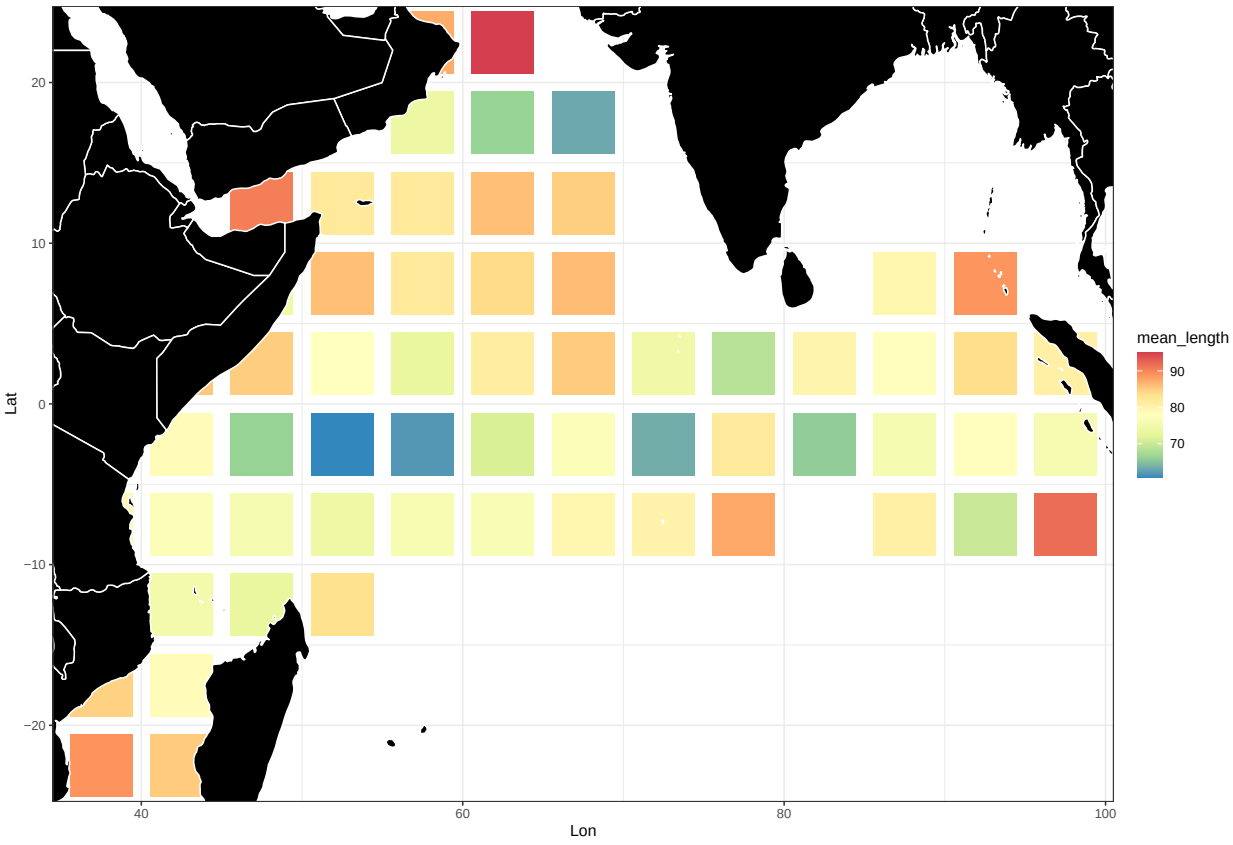
```
fcol <- 5 # the first column with LF info
lcol <- 17 # the last column with LF info
bins <- seq(30,150,10) # length of bins as a numeric vector
Nsplit <- 3 # the number of splits (the number of cells - 1)
# the directory where results will be saved
save_dir <- "D:/OneDrive - IATTC/Git/FishFreqTree/demo/"
```

Plot the length frequency data as lat-lon grids and map of mean length

```
# plot lf data as maps
make.lf.map(LF,fcol,lcol,bins,save_dir)
```




```
# plot mean length as maps
make.meanl.map(LF,fcol,lcol,bins,save_dir,s=20)
```



Find the default 4-cell combination:

```
LF_Tree <- run_regression_tree(LF,fcol,lcol,bins,Nsplit,save_dir)
```

```
##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/111/  ******
##
## [1] "Best 1st split: Lat<=-2.5"
## 9.73617287670778% variance explained

## [1] "Conditional best 2nd split is for cell 1 in split1.png: Lat<=-7.5"
## 13.829228339964% variance explained

## [1] "Conditional best 3rd split is for cell 2 in split2.png: Lon<=42.5"
## 16.1713507891749% variance explained
```

Check the default 4-cell combination results

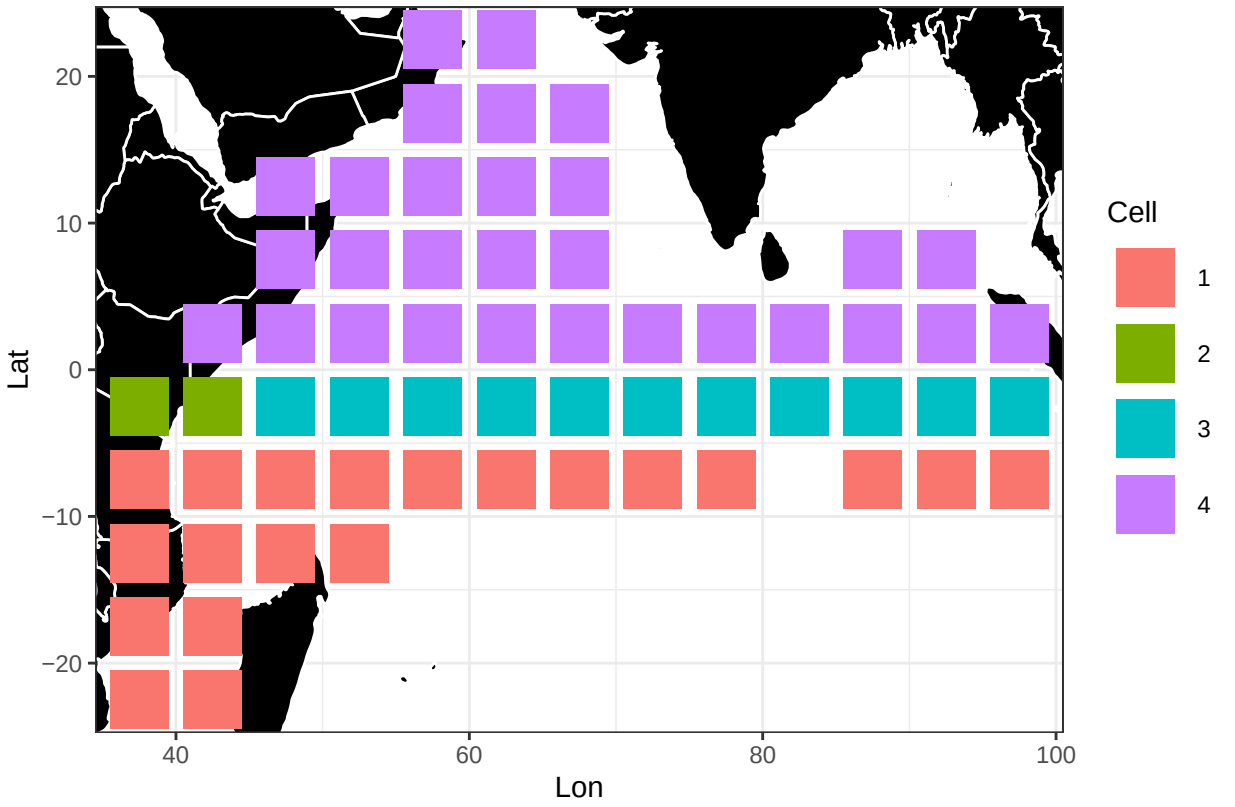
```
head(LF_Tree$LF)
```

```
## # A tibble: 6 x 21
## # Groups:   year, quarter, lat, lon [6]
##    year quarter   lat   lon    `30`   `40`   `50`   `60`   `70`   `80`   `90`
##   <dbl>   <dbl> <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1    31       1  -2.5  52.5 0.0195  0.0637 0.108  0.140  0.138  0.121  0.0908
## 2    31       1  -2.5  57.5 0.0157  0.0705 0.125  0.154  0.133  0.0966 0.107
## 3    31       1  -2.5  62.5 0.0214  0.0646 0.0986 0.143  0.134  0.116  0.0919
## 4    31       1   2.5  62.5 0.0114  0.0496 0.101  0.136  0.145  0.126  0.0966
## 5    31       1   2.5  67.5 0.00759 0.0511 0.0926 0.141  0.155  0.124  0.0926
## 6    31       2  -2.5  52.5 0.0216  0.0594 0.112  0.0971 0.0989 0.115  0.126
## # ... with 10 more variables: 100Â <dbl>, 110Â <dbl>, 120Â <dbl>, 130Â <dbl>,
## #   140Â <dbl>, 150Â <dbl>, dummyÂ <lgl>, Flag1Â <dbl>, Flag2Â <dbl>,
## #   Flag3Â <dbl>
```

```
# a summary of the three splits
LF_Tree$Record
```

```
##        Key Value Cell Var_explained
## Split1 Lat  -2.5   NA    0.09736173
## Split2 Lat  -7.5    1    0.13829228
## Split3 Lon  42.5    2    0.16171351
```

```
# map the 4 cells
make.split.map(LF_Tree$LF,Nsplit,save_dir)
```



In addition to the default combination, you can also manually explore other 4-cell combinations. For example, in the figure above cell#2 only includes two 5*5 grids, so you want to explore some other 4-cell combinations. You can run the regression tree again with the second best 2th split:

```
LF_Tree2 <- run_regression_tree(LF,fcol,lcol,bins,Nsplit,save_dir,manual = TRUE, select = c(1,2,1))
```

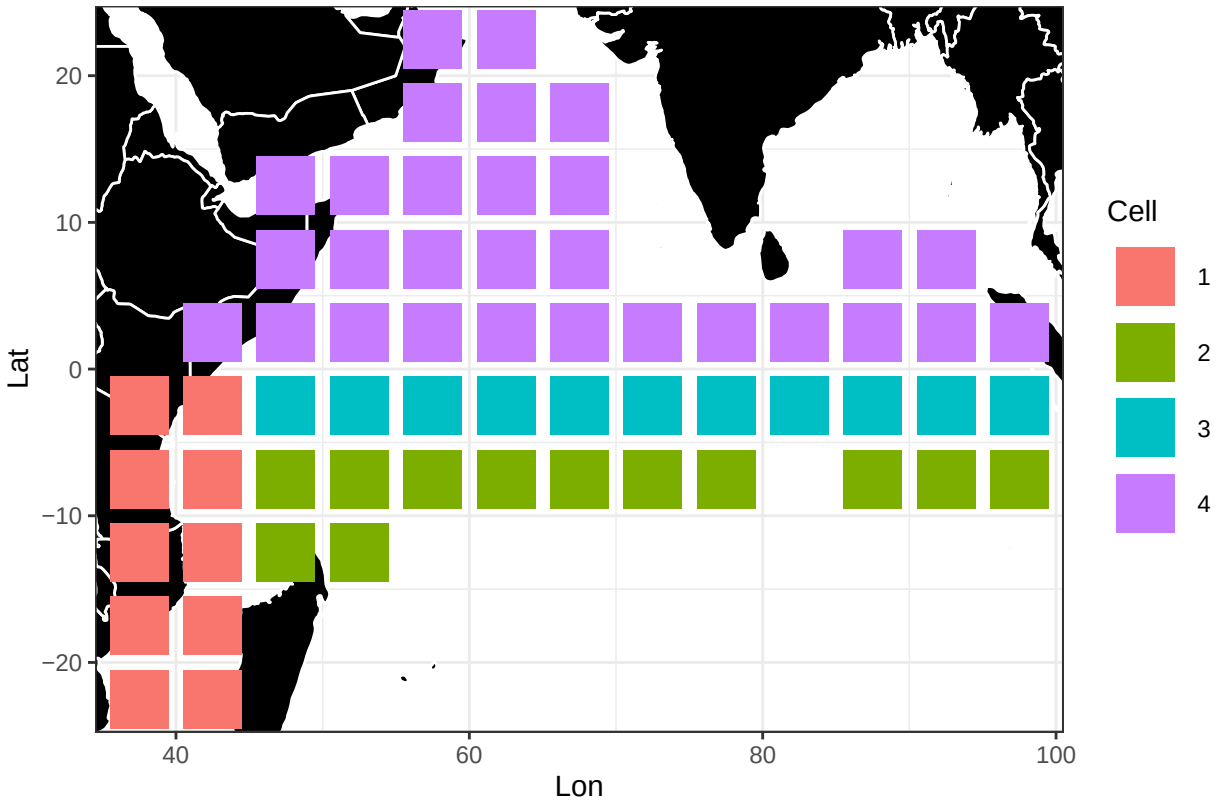
```
##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/121/  ******
##
## [1] "Best 1st split: Lat<=-2.5"
## 9.73617287670778% variance explained

## [1] "Manual 2nd split is for cell 1 in split1.png: Lon<=42.5"
## 13.2404210075085% variance explained

## [1] "Conditional best 3rd split is for cell 2 in split2.png: Lat<=-7.5"
## 16.4170943892899% variance explained
```

```
# a summary of the three splits
LF_Tree2$Record
```

```
##        Key Value Cell Var_explained
```

```
## Split1 Lat  -2.5   NA    0.09736173
## Split2 Lon  42.5    1    0.13240421
## Split3 Lat  -7.5    2    0.16417094
```

```
# map the 4 cells
make.split.map(LF_Tree2$LF,Nsplit,save_dir)
```



This setting leads to a even higher proportion of variance explained (16.4% vs. 16.2%) than the default run. Also, the number of 5*5 grids within each cell is more reasonable. In fact, you can use the loop function to explore and compare multiple 4-cell combinations:

```
loop_dir <- paste0(save_dir,"loop/")
dir.create(loop_dir)
```

```
## Warning in dir.create(loop_dir): 'D:\OneDrive -
## IATTC\Git\FishFreqTree\demo\loop' already exists
```

```
LF_Tree_Loop <- loop_regression_tree(LF,fcol,lcol,bins,Nsplit,save_dir=loop_dir,max_select = 3) #  cons
```

```
##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/loop/111/  ******
```

```
##
## [1] "Best 1st split: Lat<=-2.5"
## 9.73617287670778% variance explained

## [1] "Conditional best 2nd split is for cell 1 in split1.png: Lat<=-7.5"
## 13.829228339964% variance explained

## [1] "Conditional best 3rd split is for cell 2 in split2.png: Lon<=42.5"
## 16.1713507891749% variance explained

##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/loop/211/  ******
##
## [1] "Manual 1st split: Lat<=2.5"
## 4.66502026075569% variance explained

## [1] "Conditional best 2nd split is for cell 1 in split1.png: Lat<=-2.5"
## 10.4360136282251% variance explained

## [1] "Conditional best 3rd split is for cell 1 in split2.png: Lat<=-7.5"
## 14.5290690914814% variance explained

##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/loop/311/  ******
##
## [1] "Manual 1st split: Lon<=57.5"
## 1.37355482573027% variance explained

## [1] "Conditional best 2nd split is for cell 1 in split1.png: Lat<=-2.5"
## 8.183139459003% variance explained

## [1] "Conditional best 3rd split is for cell 1 in split2.png: Lat<=-7.5"
## 13.091845301842% variance explained

##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/loop/121/  ******
##
## [1] "Best 1st split: Lat<=-2.5"
## 9.73617287670778% variance explained

## [1] "Manual 2nd split is for cell 1 in split1.png: Lon<=42.5"
## 13.2404210075085% variance explained
```

```
## [1] "Conditional best 3rd split is for cell 2 in split2.png: Lat<=-7.5"
## 16.4170943892899% variance explained

##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/loop/131/  ******
##
## [1] "Best 1st split: Lat<=-2.5"
## 9.73617287670778% variance explained

## [1] "Manual 2nd split is for cell 1 in split1.png: Lon<=37.5"
## 11.800327867842% variance explained

## [1] "Conditional best 3rd split is for cell 2 in split2.png: Lat<=-7.5"
## 14.9531069386695% variance explained

##
##
## ***Note***: below shows the best splits in order, please check the saved figures under the directory
##
## ******  Results are saved in folder D:/OneDrive - IATTC/Git/FishFreqTree/demo/loop/111/  ******
##
## [1] "Best 1st split: Lat<=-2.5"
## 9.73617287670778% variance explained

## [1] "Conditional best 2nd split is for cell 1 in split1.png: Lat<=-7.5"
## 13.829228339964% variance explained

## [1] "Conditional best 3rd split is for cell 2 in split2.png: Lon<=42.5"
## 16.1713507891749% variance explained

##
##
## ********************************************************************************************
## Below shows the loop summary results (table saved in loop.csv)
##
## Unsorted:
##  select1 select2 select3 Var_explained
##        1       1       1     0.1617135
##        2       1       1     0.1452907
##        3       1       1     0.1309185
##        1       2       1     0.1641709
##        1       3       1     0.1495311
##
## Sorted:
##  select1 select2 select3 Var_explained
##        1       2       1     0.1641709
##        1       1       1     0.1617135
##        1       3       1     0.1495311
##        2       1       1     0.1452907
##        3       1       1     0.1309185
```

Those combinations can be compared according to the proportion of variance explained.

References

Lennert-Cody, Cleridy E., Mark N. Maunder, Alexandre Aires-da-Silva, and Mihoko Minami. 2013. "Defining Population Spatial Units: Simultaneous Analysis of Frequency Distributions and Time Series." *Fisheries Research* 139 (March): 85–92. https://doi.org/10.1016/j.fishres.2012.10.001.

Lennert-Cody, Cleridy E., Mihoko Minami, Patrick K. Tomlinson, and Mark N. Maunder. 2010. "Exploratory Analysis of Spatialtemporal Patterns in Lengthfrequency Data: An Example of Distributional Regression Trees." *Fisheries Research* 102 (3): 323–26. https://doi.org/10.1016/j.fishres.2009.11.014.