R_tools packages WKBIOPTIM3 documentation

LanAge_opt (DRAFT)

Eirini Mantzouni

Fisheries Research Institute (FRI)

Kavala, Greece

emantzo@inale.gr

1. Introduction

This tool is largely based on the tool developed by Dr Laurent Dubroca for WKBIOPTIM3 (sampalk): Simulation on the number of age readings per length class (LC) and effect on the age structure of landings (see WKBIOPTIM3 report). However, the present tool allows for optimization using actual datasets (CS and CL tables of the SDEF).

It should be noted that this approach was developed right after the WKBIOPTIM 3 meeting took place, and thus this document serves as a draft to be discussed and elaborated by the WK during the next meeting.

The *LanAge_opt* tool allows for the evaluation of the age sampling optimization effect on the age distribution of Landings for multiple years and gears, in a single area.

The required data include the consolidated CS and the CL tables. The COST methods of filling in the ALK data gaps may have been applied to the consolidated CS (optional).

The tool resamples the CA table, which is allowed to contain age data for multiple years (equal to “years_ca”). For each scenario (ie proportion in “props”), the age data of the CA table are resampled by a fraction. The resampling may be performed by length class and optionally by additional variables. The additional variables include sex, gear or time, depending on the values set for “age_by_sex”, “techStr_age’ and “timeStr_age”, respectively. Within each group, a fixed proportion (“props”) of the age data available per year is sampled. Subsequently, for every year selected in “years_AgStr”, the age structure of the landings is estimated using the actual and the reduced (resampled) age dataset.

The raising procedure relies on the methods of the COST package, but additional options are also implemented. The standard COST method applies the ALK estimated by timeStrata and by spaceStrata (ie the selected landings stratification) to all the gears.

In this tool, the following alternative approaches are possible:

a) Apply a separate ALK by gear, but common for all the timeStrata. This option should be used if the age sampling is conducted annually without further temporal stratification, and is stratified by gear,

To select this option set :

“timeStr_age” =FALSE

“techStr_age” = TRUE

b) Apply a common ALK to all gears in all timeStrata. Select this option if age data collection is implemented with no temporal or technical stratification.

To select this option set :

“timeStr_age” =FALSE

“techStr_age” = FALSE

c) Apply a separate ALK by gear, and by timeStrata.

To select this option set :

“timeStr_age” =TRUE

“techStr_age” = TRUE

d) Apply the standard COST method, whereby ALK is stratified by timeStrata only.

To select this option set :

“timeStr_age” =TRUE

“techStr_age” = FALSE

It should be noted that the tool does not allow for space stratification, and only one area may be selected and analysed each time.

The resulting age distributions of the landings for each time and technical strata defined in “strD”, obtained with the different scenarios, are plotted for visual evaluation.

In addition a number of distance metrics are estimated and plotted, to help the evaluation on the optimized sampling fraction, by quantifying the dissimilarity of the age distributions resulting from the scenarios.

The metrics include:

- EMD (Earth Mover Distance) (see documentation of the respective function: https://rdrr.io/cran/emdist/man/emd.html)
- Kolmogorov - Smirnoff distance (see documentation of the respective function: (https://www.rdocumentation.org/packages/provenance/versions/2.2/topics/KS.diss)
- Kullback- Leibler distance (see documentation of the respective function: https://www.rdocumentation.org/packages/seewave/versions/2.1.0/topics/kl.dist)
- Hellinger distance (see documentation of the respective function: https://www.rdocumentation.org/packages/statip/versions/0.2.0/topics/hellinger)

2. Libraries required

```
library(dplyr)

library(tidyr)

library(COSTcore)

library(COSTdbe)

library(sampling)

library(purrr)

library(ggplot2)

library(ggforce)

library(parallel)

library(parallelsugar)

library(statip) ## for hellinger

library(provenance) # kolmogorov-smirnoff
distlibrary(seewave) # for Kullback-Leibler distance

library(emdist) # for emd
```

In addition, a custom script to allow `mclapply` parallel computation for Windows users is also used and included in the tool (source: https://www.r-bloggers.com/implementing-mclapply-on-windows-a-primer-on-embarrassingly-parallel-computation-on-multicore-systems-with-r/).

3. Input data

The required data include the consolidated CS and the CL tables (SDEF), saved as an .Rdata file. The COST methods of filling in the ALK data gaps may have been applied to the consolidated CS (optional).

The user may either load the directly the consolidated datasets as an Rdata file or import the raw tables of the SDEF format (TR, HH, HL, SL, CA, CL) as .csv files and use the *CS_csc_data_prep_sampalc.R* script for data validation and consolidation, and exportation to an .Rdata file.

3.1 Initial values/starting values for simulations

The user needs to define the following variables in the *sampalk_CS_CL_4.R script*:

- `sppName` : scientific name of species, as it appears in the SDEF data tables
- `area` : the area of interest (limited to 1 area only)
- `years_AgStr`: year(s) for which the Age structure will be estimated
- `years_ca`: year(s) selected for the resampling of the age readings
- `GEAR`: gears (select one or more or "ALL") for the estimation of the landings age structure. These should be written as they appear in the "sel.gear.col" below
- `sel.gear.col` : The column used for the technical stratification in strD (eg "foCatEu6")
- `props`: fractions for the reduction in the number of age readings compared to the original (eg seq(0.2,0.9,by=0.1))
- `strD` : temporal and technical stratification for the Landings at age estimation (eg strIni(timeStrata="quarter", techStrata="foCatEu6")). This stratification must have been applied for the CS and CL tables consolidation.
- `techStr_age`: if the age data collection is stratified by the techStrata in strD set this to TRUE (otherwise FALSE)
- `timeStr_age`: if the age data collection is stratified by time, following the temporal stratification in strD, set this to TRUE (otherwise FALSE)
- `age_by_sex#` if age sampling is stratified also by sex, set this to TRUE (otherwise FALSE)
- `n.sim`: no. of simulations for each scenario

- `type:` If the protocol was to collect n otoliths per length class, set to "fixedK". If the protocol was to collect one otolith every n fish measured, set it to "propK"

  and finally set the path and file names for the consolidated CS and CL tables, and for the results output folder.

4 Scripts files

4.1 *CS_csc_data_prep_sampalc.R*

Optional script to import the raw tables of the SDEF format (TR, HH, HL, SL, CA, CL) as .csv files and run the data validation and consolidation, and exportation to an .Rdata file.

Alternatively, the user may load the consolidated CS and the CL tables (SDEF), saved as an .Rdata file, directly to the *sampalk_CS_CL_4.R* script.

4.2 *sampalk_CS_CL_4.R*

This is the main script of the tool where the analyses are performed. The user has to set the parameters described in the 3.1 Section.

The outputs of the script include:

- A table in .csv with the number of age readings by LC and by level of stratification selected (eg. sex, quarter)
- Figure showing Landings age distribution by gear and other levels of stratification specified (eg quarter) for the various scenarios (proportions of age readings reduction).
- Figure showing the distance statistics (normalized to the maximum) between the true (original) and the optimized age distribution of the landings plotted against the fraction of reduction in the age readings by LC, arranged by stratum (eg gear- quarter- year).
- Figure showing the distance statistics (normalized to the maximum) between the true (original) and the optimized age distribution of the landings, plotted against the fraction of reduction in the age readings by LC, arranged by statistic.

4.3 fun_emantzo2.r

Script that includes various custom functions, including:

- Estimation of the age structure for the landings, using variations of the related COST functions (*RaiseAge* () of the COSTdbe package) to allow for the various options for the stratification of the ALK, as described in the Introduction.

- Resampling age estimates from the CA table

The script is sourced within *sampalk_CS_CL_4.R* and no user action is required or advised.

4.4 mcmapply_hack.R

Script to allow for the parallelized version of `lapply`, for MS Windows users, adapted from: https://www.r-bloggers.com/implementing-mclapply-on-windows-a-primer-on-embarrassingly-parallel-computation-on-multicore-systems-with-r/.

The script is sourced within *sampalk_CS_CL_4.R* and no user action is required or advised.

4.5 Outputs interpretation (Tables, Figures, .csv files)

The outputs of the tool include:

- A table in .csv with the number of age readings by LC and by level of stratification selected (eg. sex, quarter)

- Figure showing Landings age distribution by gear and other levels of stratification specified (eg quarter) for the various scenarios (proportions of age readings reduction).

- Figure showing the distance statistics (normalized to the maximum) between the true (original) and the optimized age distribution of the landings plotted against the fraction of reduction in the age readings by LC, arranged by stratum (eg gear- quarter- year).

- Figure showing the distance statistics (normalized to the maximum) between the true (original) and the optimized age distribution of the landings, plotted against the fraction of reduction in the age readings by LC, arranged by statistic.

5. Case study example

A case study using the LanAge_opt tool for *Mullus barbatus* in GSA 22 (Aegean Sea-Greece) is presented in the WKBIOPTIM3 report.

6. Conclusions

The present tool allows investigation of the potential reduction (optimization) in the number of age readings by length class, by evaluating the effects on the age distribution of the landings. Nevertheless, age estimation is indispensable for a number of parameters (eg growth rate, life history patterns, recruitment, mortality rates), that are fundamental to fisheries dynamics and stock assessment and thus, to the sustainable management of fisheries resources. Consequently, for the safe optimization of age estimation, a broad set of factors should be considered.

7. Improvements required

The tool presented here has been inspired by the WKBIOPTIM3, but since it was developed after the meeting conclusion, it should be discussed and elaborated by the group at the – hopefully- next meeting.