BioIndex (v. 3.0) - Tutorial

March, 2021

R and Rstudio softwares

BioIndex routine (version 3.0) is developed with Rstudio running R version 4.0.4. The use of Rstudio software is strongly recommended.

Download the precompiled binary distribution of the R base system (version 4.0.4) from the **Comprehensive R Archive Network (CRAN)** https://cran.r-project.org/bin/windows/base/release.html and install it in your preferred folder. When the software is completely installed open Rstudio using the opportune R version.

In case Rstudio is not already installed, download it from the official web site: https://rstudio.com/products/rstudio/download/#download

Libraries installation

BioIndex routine needs that some supplementary libraries are installed (this step is necessary only the first time that you run the routine on your computer):

- sp
- hms
- mgcv
- rgdal
- tcltk2
- ggplot2
- gridExtra
- svDialogs
- MEDITS
- RoMEBS

In this new software version the function load_lobraries() was introduced for the first time. This function automatically checks the presence of the needed libraries in R andinstalls the missing ones. In any case the only library that needs to be installed by the user is "tcltk2". There are three possibilities to install it:

1. run commands from the console (in this case you need an internet connection):

```
install.packages("tcltk2")
```

2. install from repositories: (also in this case you need an internet connection):
   - open the menu *Packages* and then *Install package(s)*

- select the closest mirror to you position
- select the zip file(s)
- select the package(s) to be installed

3. install from zip files:

- download libraries from the **CRAN** https://cran.r-project.org/web/packages/index.html
  - open the menu *Packages* and then *Install package(s) from local files…*
  - select the zip file(s)

Initialization of the routine

The BioIndex routine is distributed as a zip file: *BioIndex_<version_number>.zip*. After you have extracted all the files from the zip in a preferred folder on your PC, you will find the main script *BioIndex_<version_number>.R* and three folders: *input*, *output*, *scripts*. Other files and folders are part of the *BioStand* routine.

The TA, TB and TC files (.csv format with semicolon as values' separator) should be placed in the respective directories of the *input* folder. In the *input* folder there are two more .csv files

- *maturity_sizes.csv* that should be compiled with the values of the cutoffs to be used for the selection of recruits (juveniles) and spawners (adults)
- GSAs_coordinates.csv containing the coordinate ranges for the study area (GSA) to be applied to resize maps plot.

The *output* folder is used by the routine to store the outputs of the analyses. At the end of each analysis it is strongly recommended to store the outputs in another folder to exclude the possibilities of overwriting the files.

The folder *scripts* contains all the functions and utilities used by the routine to perform the analyses.

Running the routine

First open the main script of the routine with the name *BioIndex_<version_number>.R* in R environment or in Rstudio.

- in R environment: open the menu *File* and then the sub-menu *Open script*. Select from the folder the file *BioIndex_<version_number>.R*.
- in Rstudio environment: open the menu *File* and then the sub-menu *Open file*. Select from the folder the file *BioIndex__<version_number>.R*.

The first part of the routine consists in cleaning objects from the workspace and installing/loading the needed "tcltk2" library.

To run the command in R, select with the mouse the row of the script with the command you want to run and press the following buttons combination: **<CTRL>+R**

if you are using **Rstudio** you need to press the following buttons combination:
<CTRL>+<ENTER>

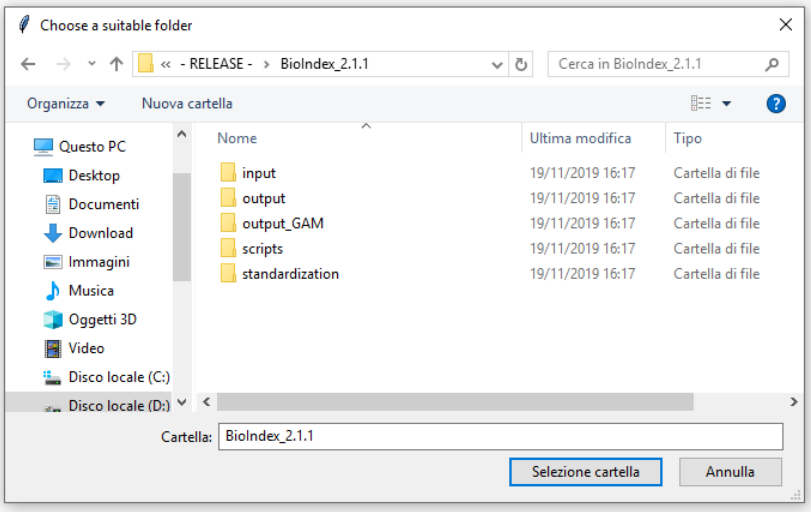
At this step you have 2 alternatives:

1) run the whole script at once clicking the button “source” in Rstudio.
2) You can run the code step by step following the instructions below described.

```
##############################################################################################
#                                                                                            #
#   BioIndex v.3.0                                                                           #
#   Developed on R 4.0.4 (x86-64bit)                                                         #
#   February 2021                                                                            #
#                                                                                            #
#   Authors:                                                                                 #
#   Walter Zupa, Loredana Casciaro, Isabella Bitetto, Maria Teresa Spedicato                 #
#   Coispa Tecnologia & Ricerca - Stazione sperimentale per lo Studio delle Risorse del Mare #
#                                                                                            #
#   In case of use of the software, the Authors should be cited.                             #
#   If you have any comments or suggestions please contact the following e-mail address:     #
#   zupa@coispa.it                                                                           #
#   BioIndex is believed to be reliable. However, we disclaim any implied warranty or        #
#   representation about its accuracy, completeness or appropriateness for any particular    #
#   purpose.                                                                                 #
##############################################################################################
```

Once all the “tcltk2”library is correctly loaded without errors you have to select the working directory. Running the following commands the routines opens a window for the assisted selection of the folder. The routine runs correctly only whether the folder containing the *BioIndex_<version_number>.R* file is selected.

```
rm(list=ls(all=TRUE))
#-------------------------------------------------------------
# Selection of the working directory
#-------------------------------------------------------------

install.packages("tcltk2")
library(tcltk2)
wd <- tk_choose.dir(getwd(), "Choose a suitable folder")
setwd(wd)
```

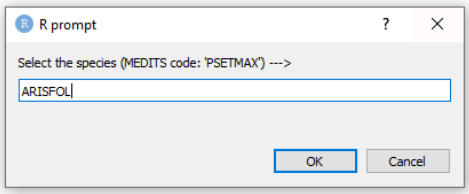


The initialization of the routine starts with the selection of the species for the analysis.

ATTENTION: Use both uppercase and lowercase letters without spaces to write the name of the species in the form of the MEDITS code, as they are reported in the MEDITS manual (AAVV, MEDITS-Handbook. Version n. 9, 2017. http://www.sibm.it/MEDITS 2011/principaledownload.htm)

Running the following code a box will pop-up for the selection of the species. "ARISFOL" is set as default value.

```
#-----------------------------------------------------------------
# initialization
#-----------------------------------------------------------------
source(paste(wd, "/scripts/01_intro.r", sep=""))
```



For the analysis the routine uses meta-database files that have origin from the merge of the TA file respectively with TB and TC ones. The routine will produce automatically two tables containing the merge results that will be saved in the output folder as .csv files:

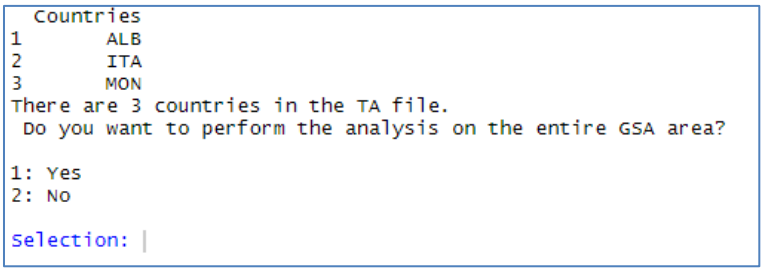
- *mergeTATB_<species_name>.csv*
- *mergeTATC_<species_name>.csv*

```
########################
metaDB preparation
########################

TA file correctly read
TB file correctly read
```

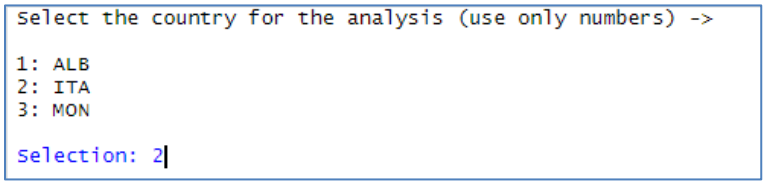
```
TC file correctly read

Merging TA-TB files
TA-TB files correctly merged
Merge TA-TB files saved in the following folder: 'D:/Documents and Settings/Utente
/Documenti/GitHub/BioIndex/R_BioIndex_3.0/output/mergeTATB_ARISFOL.csv
'
Merging TA-TC files
TA-TC files correctly merged
Merge TA-TC files saved in the following folder: 'D:/Documents and Settings/Utente
/Documenti/GitHub/BioIndex/R_BioIndex_3.0/output/mergeTATC_ARISFOL.csv'
```

Since version 3.0 BioIndex includes also some formal checks for TA, TB, TC table formats that are derived from RoMEBS R library. The results of the check will be reported in the "Logfile" directory included in the "output" folder. In case of errors in the TX tables, correct data and run again the script.

In case the study area is a GSA in which 2 or more countries are included, the code allows to perform the analysis at GSA level or at country level. Hence, the user is asked to answer to the following question:



In case the choice is 2 (NO), the user is asked to select the reference country:



In the next step e check of the survey is performed assessing whether the following condition:

- Was the same gear used along the years?
- Was the same vessel used along the years?
- Was the same number of hauls used along the years?
- Were hauls allocated in the same position along the years?

The first three checks were done by the routine:

```
#########################
Check of survey data
#########################
During the survey a different number of hauls per year was used. The standardizati
on of the indices with GLM/GAM models is suggested.
```

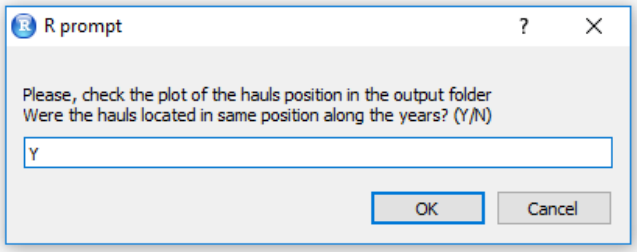
After, a plot of the hauls position is generated to allow the qualitative check of the hauls allocation during the years. Hence, the user is asked to answer to the following question

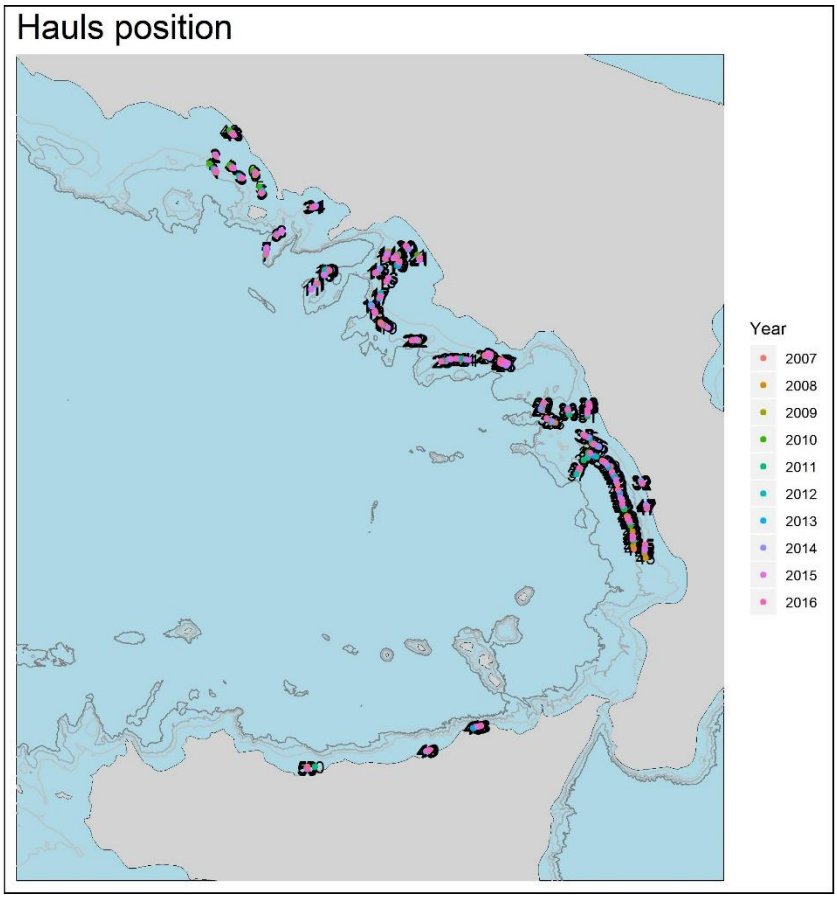
"Where the hauls located in the same position along the years?"

ATTENTION: check that the coordinate range of the GSA is uploaded in the file "~/input/GSAs_coordinates.csv"

An example of the check results is reported hereafter:

```
Please, wait a moment. Elaboration of the hauls plot in progress
Regions defined for each Polygons
Regions defined for each Polygons
Bubble plot of Hauls position correctly saved

Please, check the plot of the hauls position in the output folder

 1 check/s over 4 indicate/s that the standardization of the indices with GLM/GAM
models is reccomended
```





In case the checks have negative answers, the user is addressed toward the standardization of the indices with GLM/GAM models.

Some analyses included in BioIndex use the 30" GFCM geographical grid. Hence, catch data and biological data are merged with the grid producing two meta-DB files:

- *<species_name> - allGSAs_metaDB_catch in GRID.csv*
- *<species_name> - allGSAs_metaDB_biological in GRID.csv*

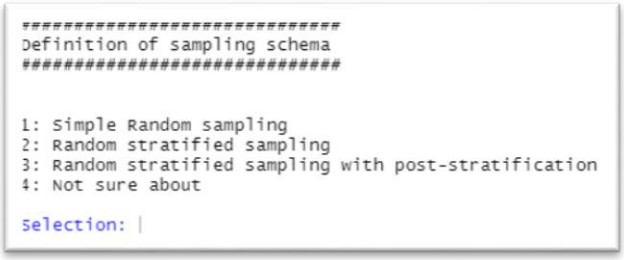
```
########################
spatial metaDB preparation
########################
Catch metaDB saved in the following folder: 'D:/Documents and Settings/Utente/Docu
menti/GitHub/BioIndex/R_BioIndex_3.0/output/ARISFOL - allGSAs_metaDB_catch in GRID
.csv

Biological metaDB saved in the following folder: 'D:/Documents and Settings/Utente
/Documenti/GitHub/BioIndex/R_BioIndex_3.0/output/ARISFOL - allGSAs_metaDB_biologic
al in GRID.csv
```

The second part of the software is dedicated to data analysis. In this software section the user will be guided in the use of the software by pop-up messages.

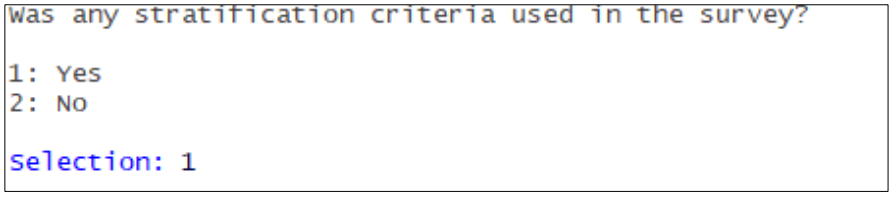
To start data analysis run the following part of the code:

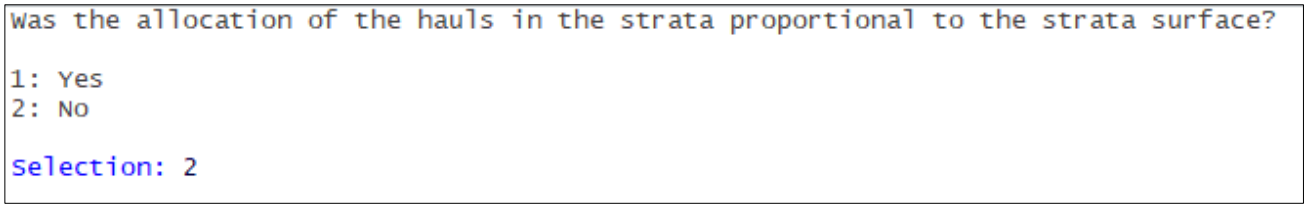
```
#-------------------------------------------------------------
# Data analysis
#-------------------------------------------------------------
source(paste(wd, "/scripts/02_Data_analysis.r", sep=""))
```

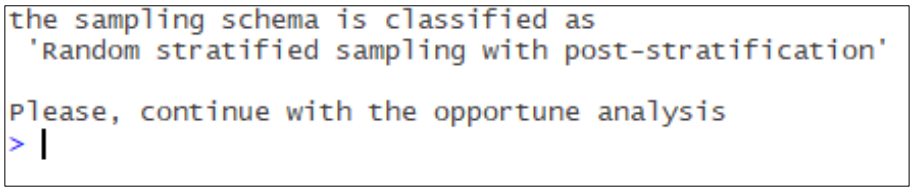
Once the code is run, the user is asked to define the sampling schema adopted during the survey. Three possibilities are available:



In case the user is not sure of the right choice, select the choice number 4 to be guided in the selection by a brief questionnaire to categorize the sampling protocol adopted, in order to continue the estimation of the indices using the opportune formulas.







At this point the user should continue running the part of the script corresponding to the selected sampling protocol:

```
the sampling schema is classified as:
 'Random stratified sampling'

Please, continue with the opportune analysis
############################
Random stratified sampling
############################
```

The routine continues estimating the time series of abundance (n/km^2) and biomass (kg/km^2) indices.

A pops-up box will appear for the selection of the depth range to be applied to the analysis. The box will not appear in case of the "Simple Random Sampling"

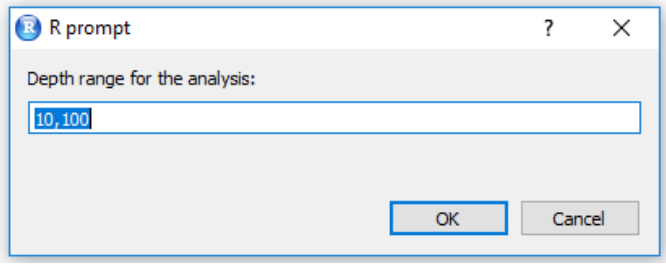
To estimate the indices time series in the case of Random Stratified (RSS) and Random Penalized Sampling (RPS, Random stratified sampling with post-stratification) the user have to define the stratification scheme adopted during the survey. BioIndex is able to use up to 6 strata defined by the user.

ATTENTION: check that the strata information were updated in the "~\scripts\utilities\strata.csv" file.

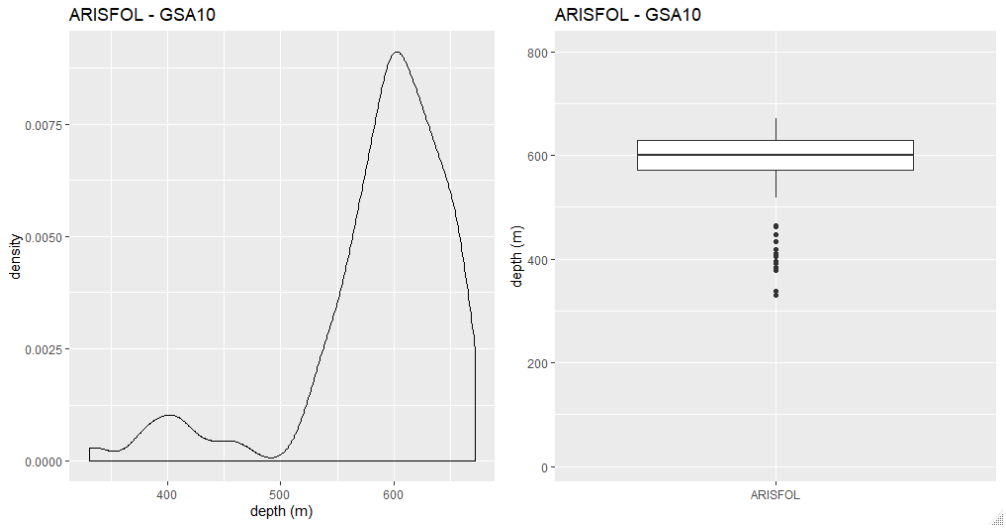
GSA	COUNTRY	CODE	MIN_DEPTH	MAX_DEPTH
10	ITA	1	10	50
10	ITA	2	50	100
10	ITA	3	100	200
10	ITA	4	200	500
10	ITA	5	500	800
…	…	…	…	…

The estimation of the indices could be performed using all the different assemblage of the contiguous strata, defining the minimum and maximum values of the depth range to be used:

ATTENTION: separate the values only with comma.



The extension of the depth range for the studied species could be assessed using the plots saved in the output folder showing the occurrence frequency and a box plot of positive hauls in the 0-800m depth range.

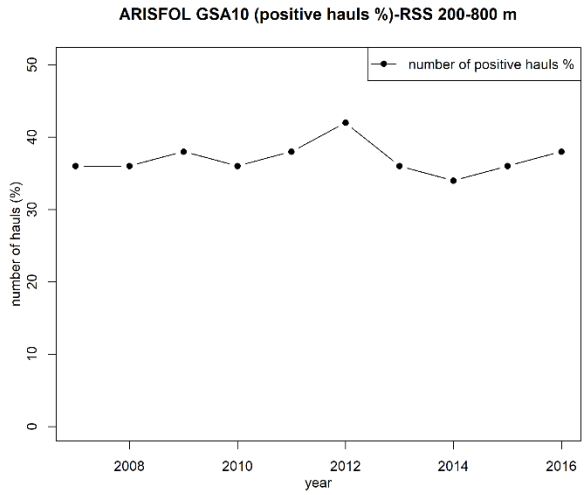


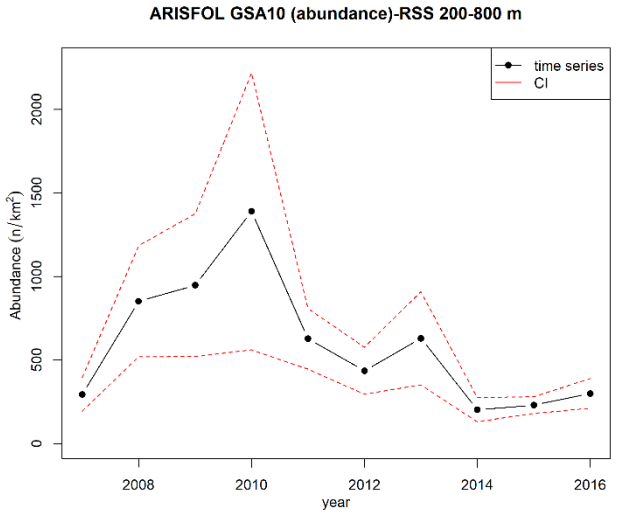
The following files are saved in the output folder:

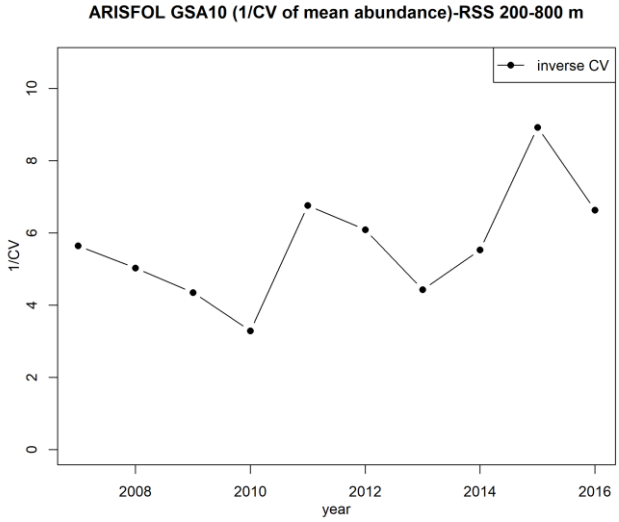
- *depth.distribution_(boxplot)ARISFOL_GSA10.jpg*
- *depth.distribution_ARISFOL _GSA10.jpg*

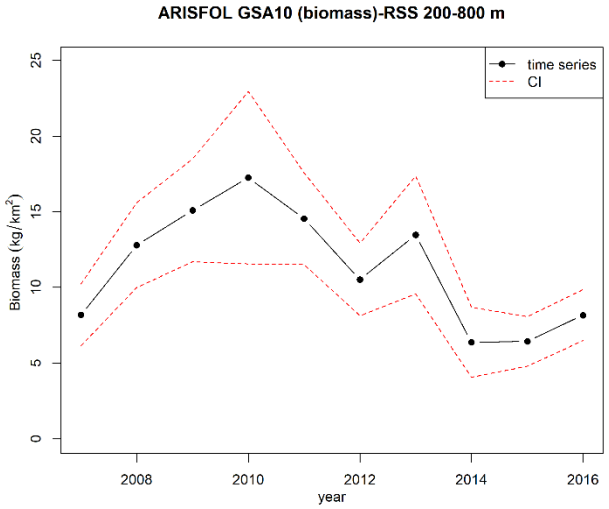
Moreover, the following .csv and .tiff files of the estimated indices are saved:

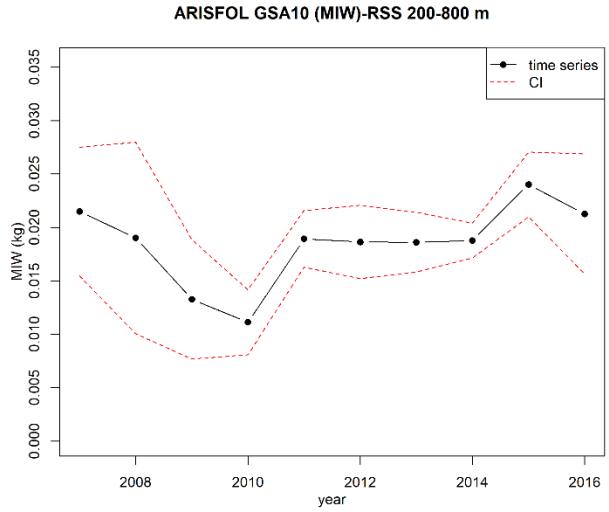
- *ARISFOL_GSA10_(positive hauls)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*
- *ARISFOL_GSA10_(abundance)-Random_Stratified_Sampling_200-800 m_Timeseries.csv*
- *ARISFOL_GSA10_(abundance)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*
- *ARISFOL_GSA10_(abundance)-FEMALES-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*
- *ARISFOL_GSA10_(abundance)-FEMALES-Random_Stratified_Sampling_200-800 m_Timeseries.csv*
- *ARISFOL_GSA10_(abundance)-MALES-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*
- *ARISFOL_GSA10_(abundance)-MALES-Random_Stratified_Sampling_200-800 m_Timeseries.csv*
- *ARISFOL_GSA10_(inverseCV of abundance)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*
- *ARISFOL_GSA10_(biomass)-Random_Stratified_Sampling_200-800 m_Timeseries.csv*
- *ARISFOL_GSA10_(biomass)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*
- *ARISFOL_GSA10_(MIW)-Random_Stratified_Sampling_200-800_Timeseries.csv*
- *ARISFOL_GSA10_(MIW)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*




ARISFOL GSA10 (abundance)-RSS 200-800 m
time series
CI
Abundance (n/km²)
0
500
1000
1500
2000
2008
2010
2012
2014
2016
year

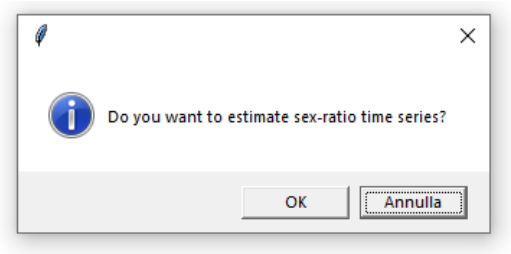

ARISFOL GSA10 (1/CV of mean abundance)-RSS 200-800 m
inverse CV
1/CV
0
2
4
6
8
10
2008
2010
2012
2014
2016
year






```
#########################
Time series of indices
#########################

[1] "Select the depth range for the analysis"

 Estimation of abundance indices completed

 Estimation of biomass indices completed

 Estimation of MIW completed
Time series of indices - completed
```

To estimate the time series of the species **sex ratio** in the selected GSA select “OK” on the following box message:

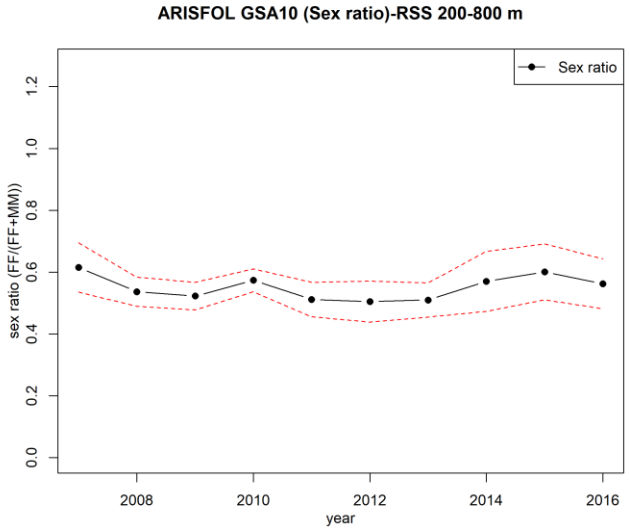


The routine uses the depth range previously selected for the abundance and density indices.

```
########################
Sex-ratio time series
########################

Sex-ratio analysis - completed
```

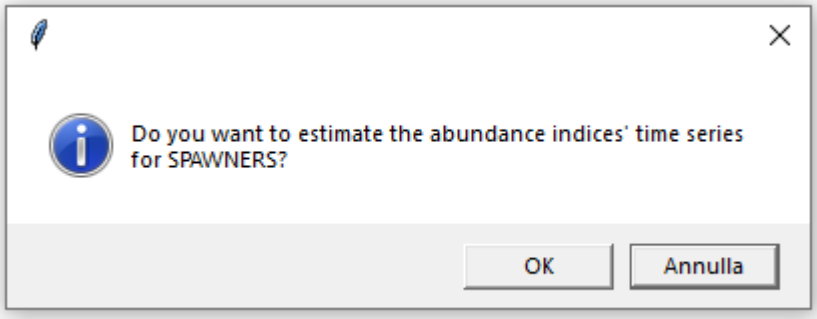
The output are the following .csv and .tiff files:

- *ARISFOL_GSA10_(Sex ratio)-Random_Stratified_Sampling_200-800 m_Timeseries.csv*
- *ARISFOL_GSA10_(Sex ratio)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*



The **spawners** are selected using the threshold value inserted in the *maturity_sizes.csv* file ("~\input\maturity_sizes.csv"). Before continue, check that the ***maturity_sizes.csv*** file is updated.

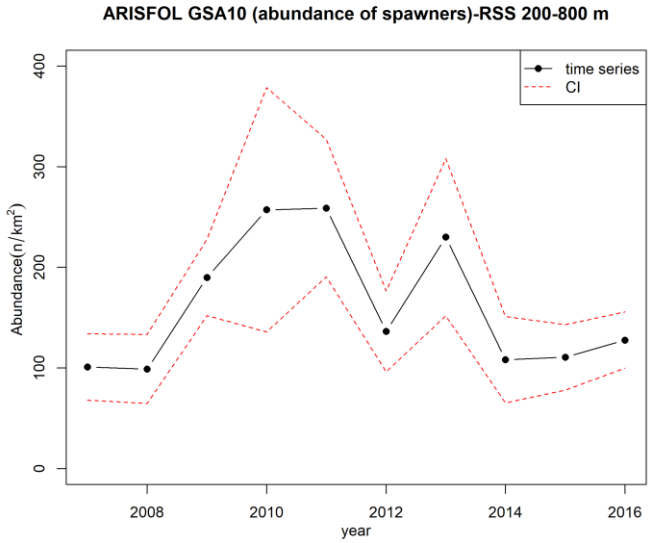
ATTENTION: Only female specimens are selected for the estimation of the spawners indices.




```
#------------------------------------------------------------
# Spawners' abundance indices
#------------------------------------------------------------

Spawners' indices analysis - completed
```

The output are the following .csv and .tiff files:

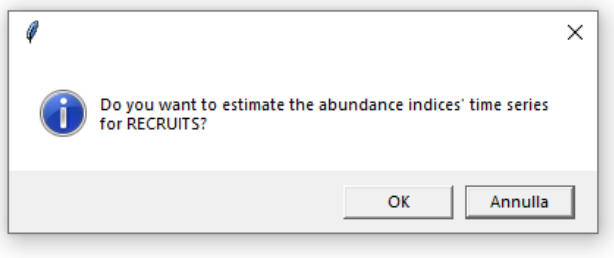
- *ARISFOL_GSA10_(abundance of spawners)-Random_Stratified_Sampling_200-800 m_Timeseries.csv*
- *ARISFOL_GSA10_(abundance of spawners)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*



In case an error is generated by the lack of threshold values the analysis is automatically skipped showing the following message:

"Spawners' indices analysis skipped - (Run Error)"

The **recruits** are selected using the threshold value inserted in the *maturity_sizes.csv* file ("~\input\maturity_sizes.csv"). Before going ahead with the analysis **check** if you have updated the ***maturity_sizes.csv*** file.




```
#------------------------------------------------------------
# Recruits' abundance indices
#------------------------------------------------------------

Recruits' indices analysis - completed
```

The outputs are the following .csv and .tiff files:

- *ARISFOL_GSA10_(abundance of recruits)-Random_Stratified_Sampling_200-800 m_Timeseries.csv*
- *ARISFOL_GSA10_(abundance of recruits)-Random_Stratified_Sampling_200-800 m_Timeseries.tiff*

In case an error is generated by missing threshold values the analysis is automatically skipped showing the following message:

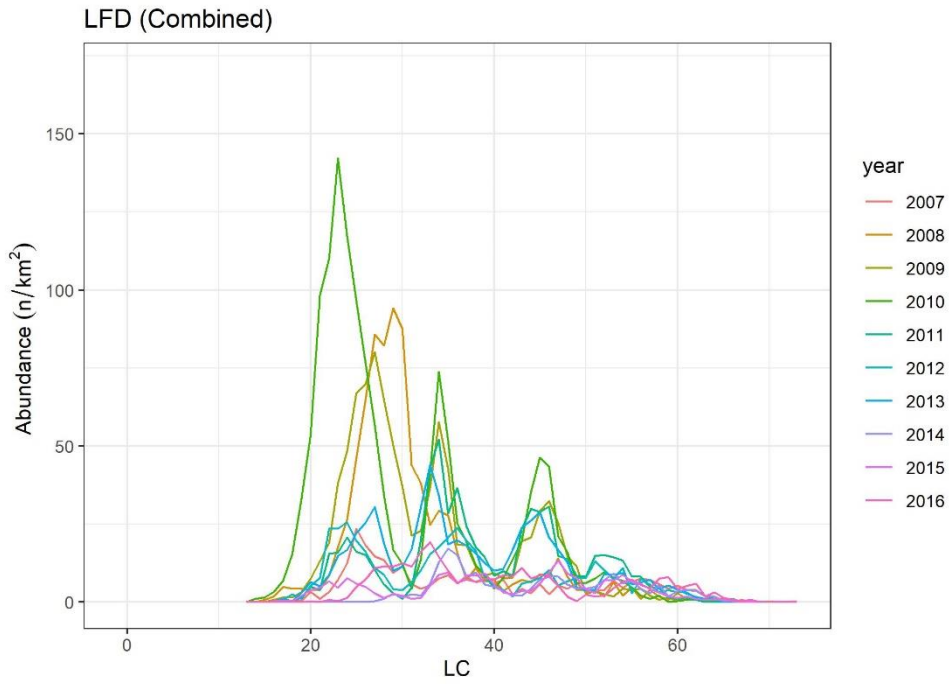
"Recruits' indices analysis skipped - (Run Error)"

To estimate the **Length Frequency Distribution** (LFD), the **median length** (50^{th}) and the **length at 95^{th} percentile** click "OK" on the following message box:

```
#------------------------------------------------------------
# LFD, L0.50 & L0.95
#------------------------------------------------------------
LFD & L0.95 analysis - completed
```

The LFDs are estimated per sex and the following csv and jpg files are saved:

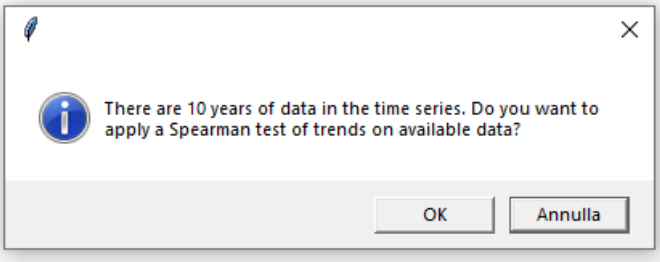
- *ARISFOL_GSA10_LFD_(Combined)_RSS.csv*
- *ARISFOL_GSA10_LFD_(Females)_RSS.csv*
- *ARISFOL_GSA10_LFD_(Males)_RSS.csv*
- *LFD_(Combined)_ARISFOL_GSA10_RSS.jpg*
- *LFD_(Females)_ARISFOL_GSA10_RSS.jpg*
- *LFD_(Males)_ARISFOL_GSA10_RSS.jpg*



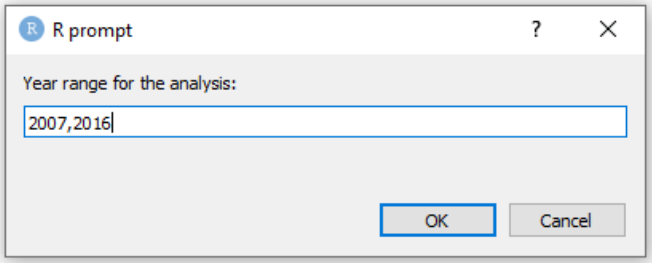
When all the LFD are computed, the LFD of sexes combined is used to estimate theL0.50 and L0.95

The results are reported in a table saved in a csv file: *ARISFOL_GSA10_L50_L95_RSS.csv.*

The **analysis of trend** could be carried out with two different tests. The first is the Spearman's test. If the time series is shorter than three years the analysis is automatically skipped. If the time series has 3 or more years of data the user is asked to continue the analysis with the available data clicking "OK" on the following box message:



The temporal range should be defined by the user. A pop-up box appears for choosing the extension of the time series to be used in the analysis of trends. Separate values with the comma. The default value is "2006,2012"

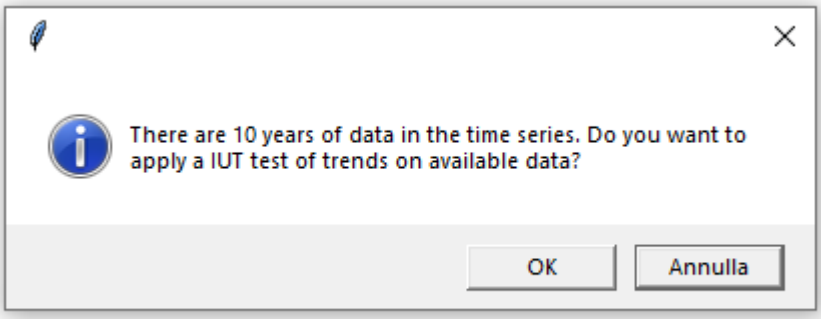



```
#-------------------------------------------------------------
# Spearman test of trends on short timeseries
#-------------------------------------------------------------

[1] "Select the year range for the analysis"
      index          r         t         p
1 abundance -0.5030303 -1.646229 0.1383337
2   biomass -0.5393939 -1.811805 0.1075932
```

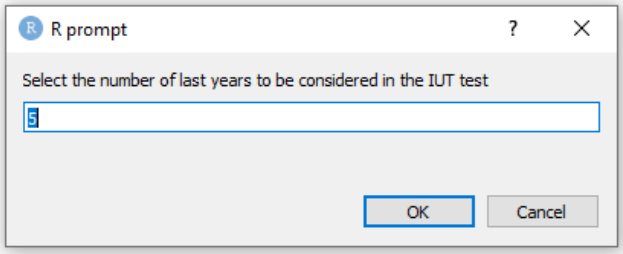
The results of the test are saved in a csv file (*ARISFOL - Spearman summary_RPS.csv*) and visualized in the console as follows:

```
[1] "Select the year range for the analysis"
      index          r         t         p
1 abundance -0.5030303 -1.646229 0.1383337
2   biomass -0.5393939 -1.811805 0.1075932
```

A second test that could be applied to the time series is the Intersection Union Test. The analysis is automatically skipped if the time series is shorter than 5 years. If the time series has 5 or more years of data the user is asked to continue the analysis with the available data clicking "OK" on the following box message:



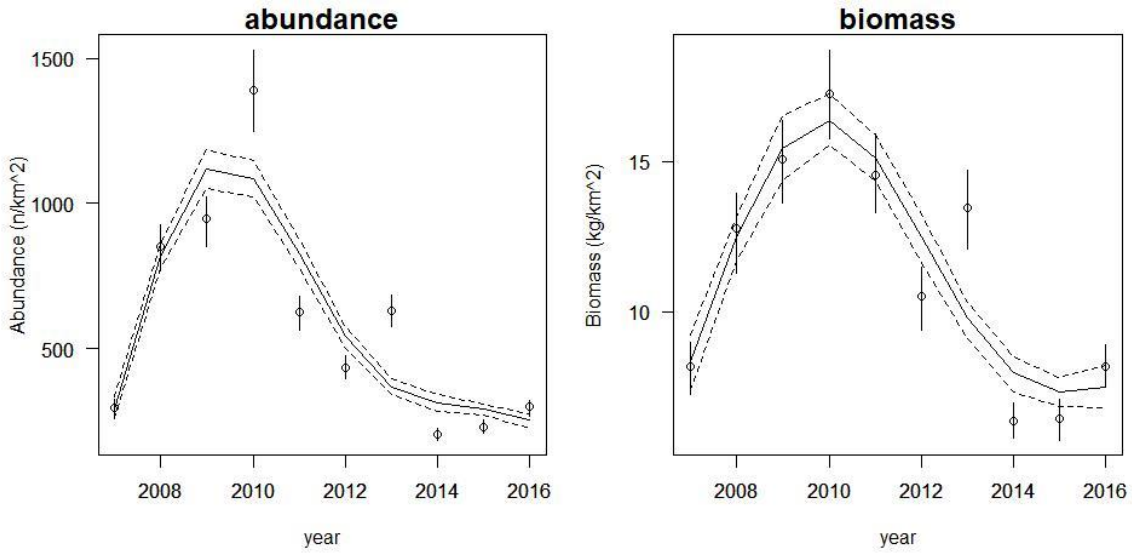
The *lastn* parameter should be set (the default value is 5). This parameter is used by the routine to set the number of last years to be considered during the analysis.



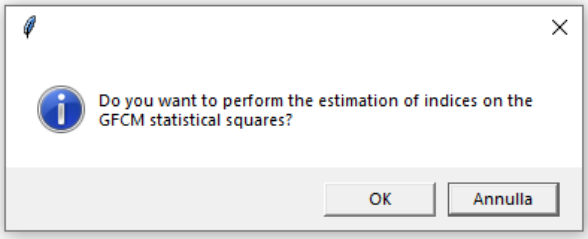

```
#-------------------------------------------------------------
# Intersection Union Test
#-------------------------------------------------------------

IUT test - completed
```

An intermediate file is generated by the routine (*GSA10IUtest5_RSS.csv*) while the results are stored in the following files:

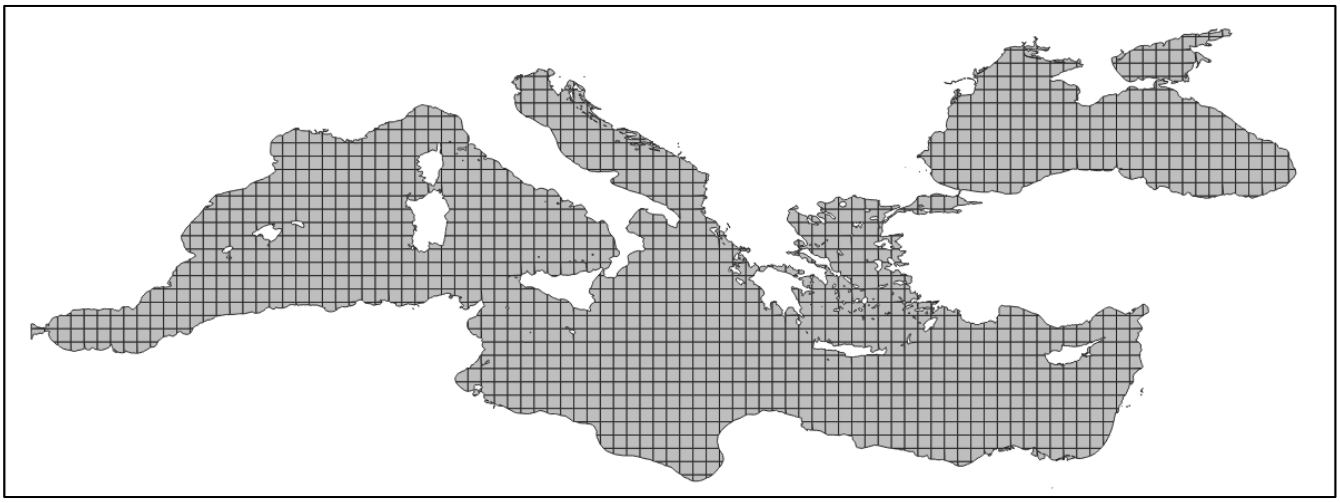
- ARISFOL_GSA10_SmoothedIndicators_RSS.jpg
- ARISFOL_GSA10_IUT_results_5years_RSS.csv



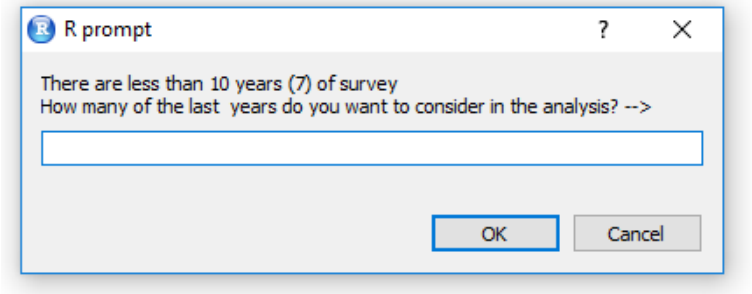
The **spatial indicators** are estimated with the resolution of the 30" GFCM grid considering only the last 10 years of the time series. The user is asked to continue the analysis or skip it.



ATTENTION: check whether the coordinate range of the GSA is uploaded in the file "~/input/GSAs_coordinates.csv"



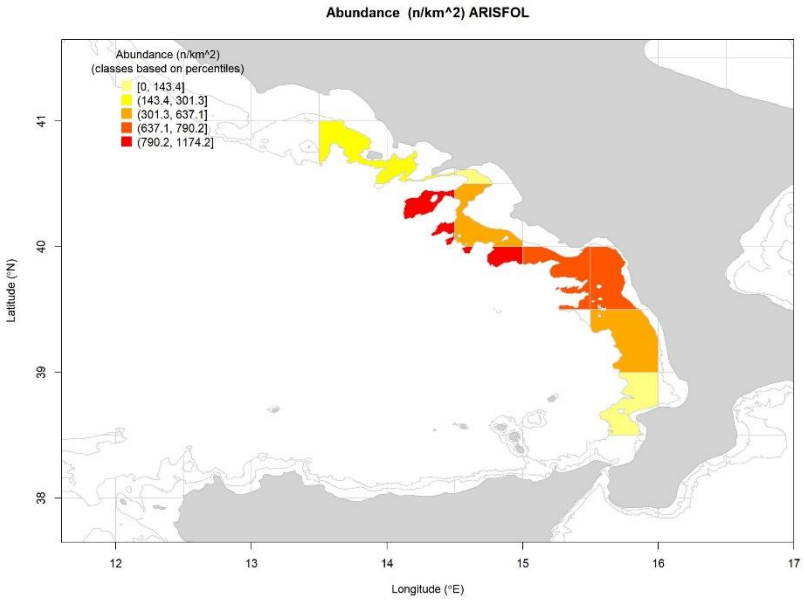
In case there are less than 10 years of survey data a box will pop-up to insert the number of the last years to be considered in the analysis.

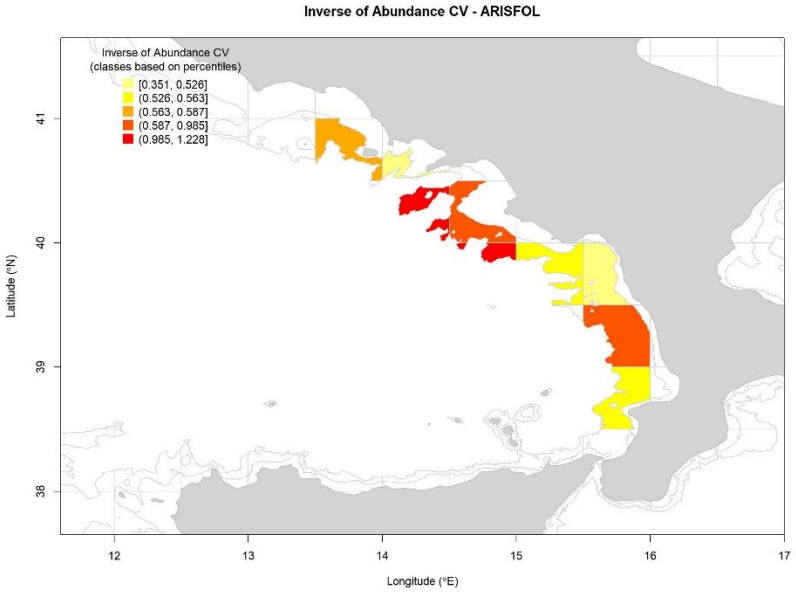


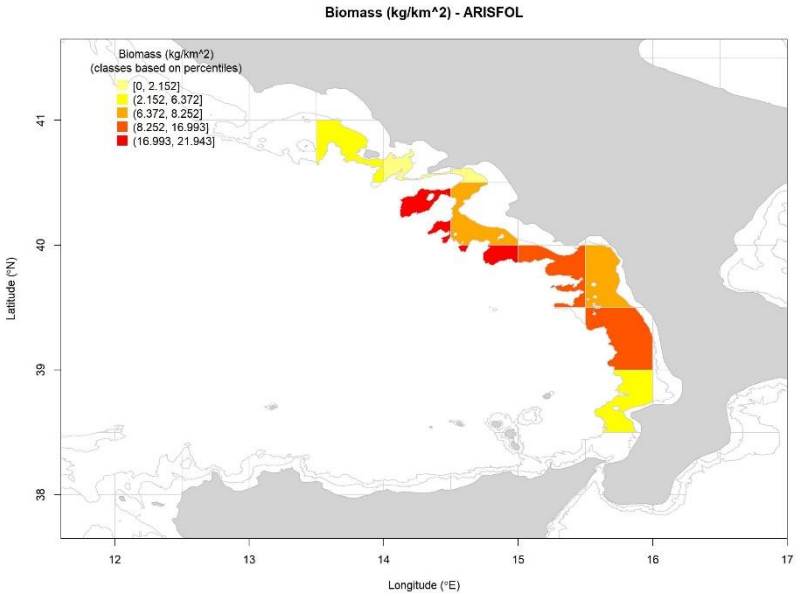

```
##############################
# Spatial analysis on GFCM grid
##############################
Abundance indices for statistical squares correctly estimated
inverse of CV of abundance indices for statistical squares correctly estimated
file of abundance indices for statistical squares saved in the following folder: '
D:/Documents and Settings/Utente/Documenti/GitHub/BioIndex/R_BioIndex_3.0/output/A
RISFOL - GFCM GRID ABUNDANCE.csv
Biomass indices for statistical squares correctly estimated
file of Biomass indices for statistical squares saved in the following folder: 'D:
/Documents and Settings/Utente/Documenti/GitHub/BioIndex/R_BioIndex_3.0/output/ARI
SFOL - GFCM GRID BIOMASS.csv
MIW for statistical squares correctly estimated
inverse of CV of MIW for statistical squares correctly estimated
file of MIW for statistical squares saved in the following folder: 'D:/Documents a
nd Settings/Utente/Documenti/GitHub/BioIndex/R_BioIndex_3.0/output/ARISFOL - GFCM
GRID MIW.csv
```

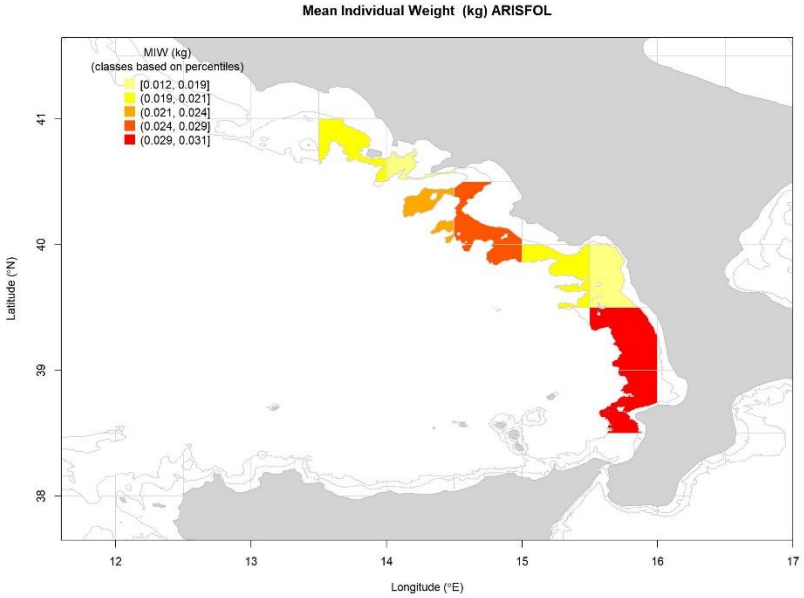
The results are reported in the following .csv and .jpg files:

- ARISFOL - GFCM GRID ABUNDANCE.csv
- ARISFOL - GFCM GRID BIOMASS.csv
- ARISFOL - GFCM GRID MIW.csv
- ARISFOL - GFCM GRID ABUNDANCE.jpg
- ARISFOL - GFCM GRID ABUNDANCE Inverse CV.jpg
- ARISFOL - GFCM GRID BIOMASS.jpg
- ARISFOL - GFCM GRID MIW.jpg

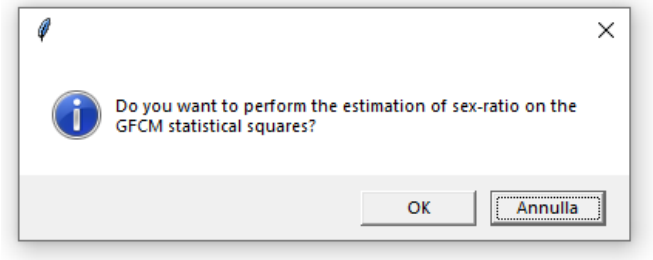




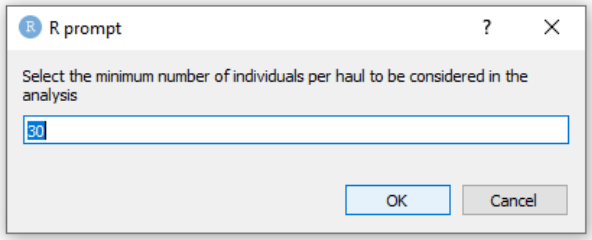




Also the **sex ratio** could be estimated over the GFCM grid. The user is asked to continue with the analysis or to skip it:

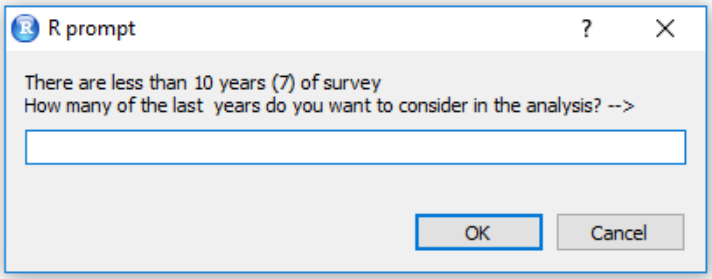


ATTENTION: set the threshold of the minimum number of individuals per haul to be considered in the analysis.



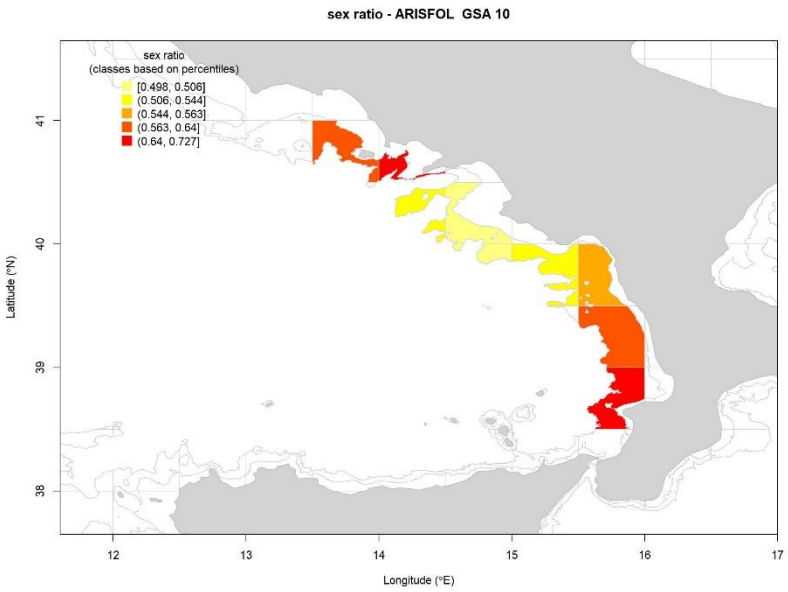

```
##############################
Sex-ratio on GFCM grid
##############################

Sex-ratio on GFCM grid - completed
```

In case there are less than 10 years of survey data a box will pop-up to insert the number of the last years to be considered in the analysis.



The outputs are stored in the following .csv and .jpg files:

- ARISFOL - GFCM SEX RATIO.csv
- ARISFOL - GFCM GRID Sex Ratio.jpg

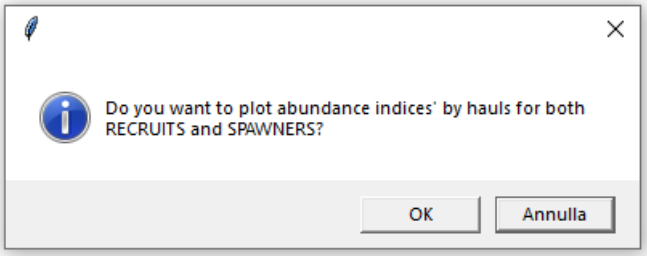


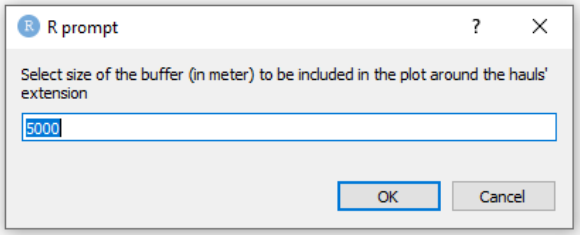
Finally, the abundance indices by hauls of the whole time series can be plotted over a map for both recruits and spawners (only female specimens).
The recruits and spawners are selected using the threshold values inserted in the *maturity_sizes.csv* file (input folder). Before running the code **check** that you have updated the ***maturity_sizes.csv*** file.

ATTENTION: check that the coordinate range of the GSA is uploaded in the file "~/input/GSAs_coordinates.csv"

The "GSAs_coordinates.csv" file contains both the coordinates range of the GSA and the preference about the bathymetrical lines to be plotted in the graphs as grey lines. Three reference bathymetrical lines could be selected respectively in the fields "depth1", "depth2" and "depth3" selecting the values in the following list: 15, 25, 35, 50, 100, 200, 500 and 800.

A	B	C	D	E	F	G	H	I	J
GSA	COUNTRY	xmin	xmax	ymin	ymax	depth1	depth2	depth3	notes
10	ITA	12.2	16.7	37.7	41.2	200	500	800	GSA10
18	ITA	15.5	20	39.8	42.5	200	500	800	GSA18
19	ITA	14.8	19.5	35.7	41	200	500	800	GSA19

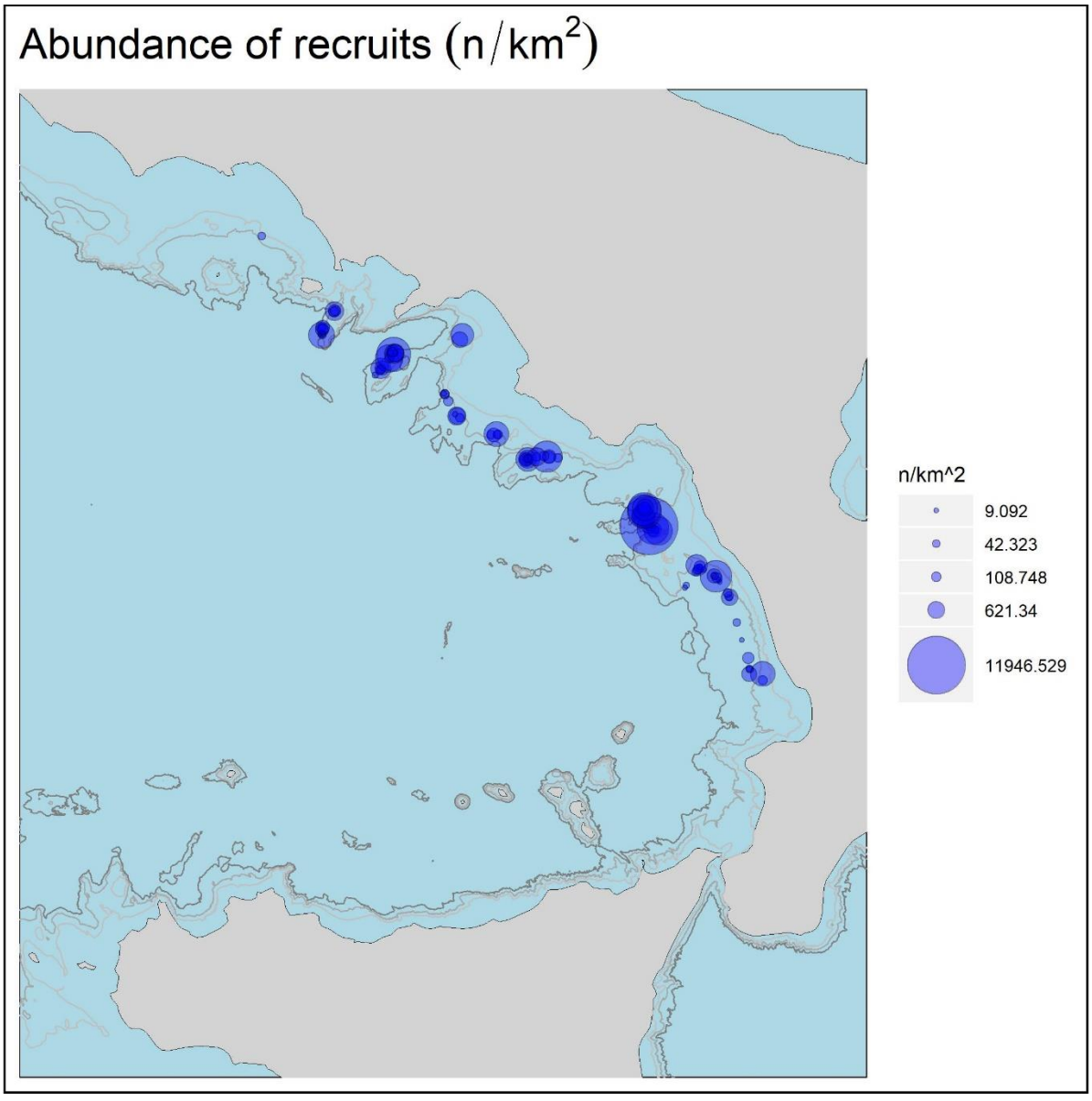





```
################################################
Bubble plots - indices of recruits and spawners
################################################

#------> check the threshold in the file "~/input/maturity_sizes.csv"
source(paste(wd, "/scripts/Bubble_plot_by_haul.r", sep=""))

## Regions defined for each Polygons
## Regions defined for each Polygons
## Bubble plot of recruits correctly saved
## Regions defined for each Polygons
## Regions defined for each Polygons
## Bubble plot of spawners correctly saved
```

The following jpg files are saved:

- ARISFOL_GSA10 -indices of RECRUITS.jpg
- ARISFOL_GSA10 -indices of SPAWNERS.jpg


Abundance of recruits (n/km^2)
n/km^2
9.092
42.323
108.748
621.34
11946.529