Index standardization - Tutorial

Walter Zupa, Isabella Bitetto, Maria Teresa Spedicato

Coispa Tecnologia & Ricerca - Stazione sperimentale per lo Studio delle Risorse del Mare

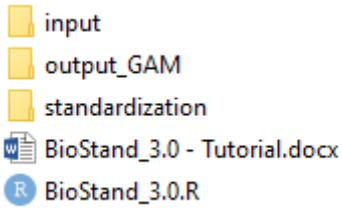
March, 2022

doi: 10.5281/zenodo.6392940

BioStand routine (version 3.0) was developed with Rstudio running R version 4.1.2. The routine runs in its own working directory and do not needs more *BioIndex* routine to produce the time series of the abundance and biomass indices.
BioStand routine uses the following supplementary libraries:

- fmsb
- Hmisc
- hms
- mgcv
- raster
- rgdal
- sp
- svDialogs
- tcltk

The routine is composed by the file *BioStand_3.0.R*, the tutorial and two additional folders ("*output_GAM"* and "*standardization"*).



Running the routine

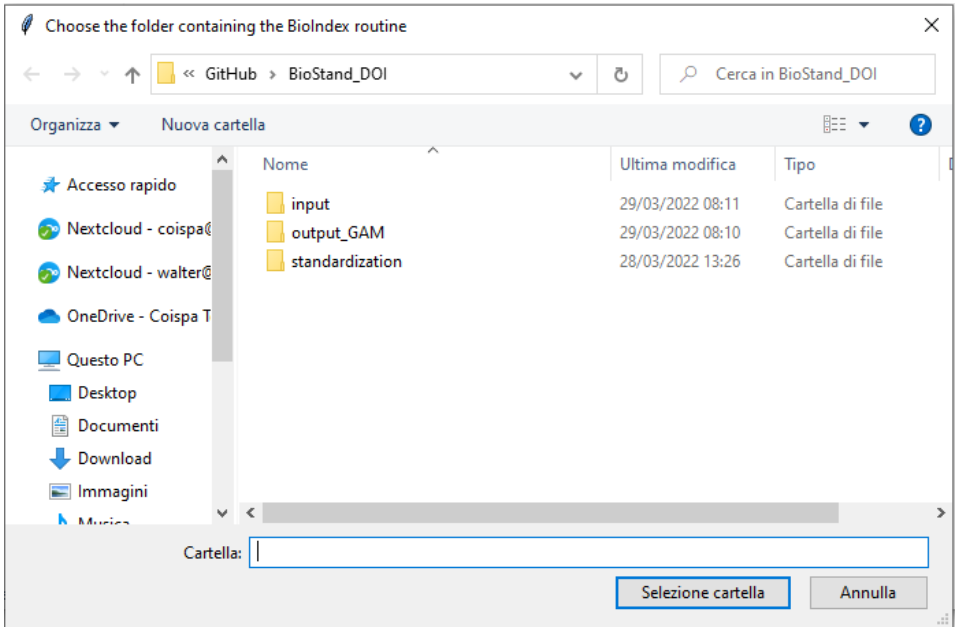
First open in R environment or in Rstudio the main script of the routine with the name *BioStand_3.0.R*.

- in R environment: open the menu *File* and then the sub-menu *Open script*. Select from the folder the file *BioStand_3.0.R*.
- in Rstudio environment: open the menu *File* and then the sub-menu *Open file*. Select from the folder the file *BioStand_3.0.R*.

The first part of the routine consists in loading the previously installed libraries. Run the following code:

```
rm(list=ls(all=TRUE))
```

```
library(sp)
library(raster)
library(rgdal)
library(hms)
library(mgcv)
library(fmsb)
library(Hmisc)
library(tcltk)
library(svDialogs)
#------------------------------------------------------------
# Selection of the working directory
#------------------------------------------------------------
wd <- tk_choose.dir(getwd(), "Choose the folder containing the BioIndex routine
")
setwd(wd)
```



The following instruction is used to read all the functions needed by the routine. Run the following code:

```
source(paste(wd,"/standardization/functions.r", sep=""))
```

The routine generate a new merge table between TA and TB files. The source files are the same used by the *BioIndex* routine, placed in the *input* folder. The merge file is different from the one used in the *BioIndex* routine. Once the merge table is generated it is saved as csv file in the folder *output_GAM* that is located in the working directory.

```
merge_TATB <- merge_TATB_function()
```

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:

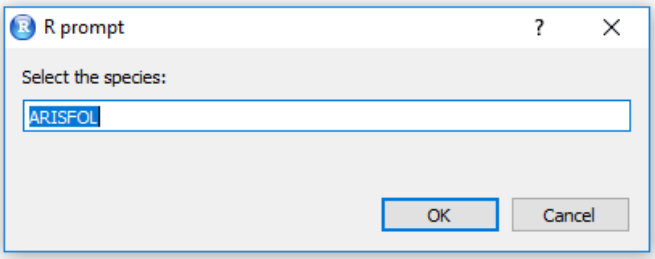
```
  Countries
1       ALB
2       ITA
3       MON
There are 3 countries in the TA file.
 Do you want to perform the analysis on the entire GSA area?

1: Yes
2: No

Selection:
```

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

```
Select the country for the analysis (use only numbers) ->

1: ALB
2: ITA
3: MON

Selection: 2
```

After a box will pop-up for the selection of the species. “ARISFOL” is set as default value.

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).




```
#-------------------------------------------------------------
# Preparation of merge TA-TB file
#-------------------------------------------------------------
merge_TATB <- merge_TATB_function()
## Merging TA-TB files
## TA-TB files correctly merged
## Merge TA-TB files saved in the following folder: 'D:/R_BioIndex_3.0/output _
GAM/mergeTATB_ARISFOL_standardization.csv'
```

The routine uses two different methods to estimate the standardized time series of the indices. The first method uses the final model to predict the value of the index to be standardized from the sampling points (locations of survey hauls), while the second method uses a grid made of points regularly distributed in the space and at a constant distance in longitude and latitude to predict the results from the final model. The user could choose between two different grid resolution: 0.003 degrees (approximately 1.8NM on the latitude) and 0.0625 degrees (the latter corresponds to the statistical grid used by Copernicus layers, approximately 3.7NM on the latitude). The grid is

filtered to select only the points belonging to the study area: in case the analysis is performed at country level, the grid is further filtered for the selected country. Run the following code:

```
#------------------------------------------------------------
# grid loading
#------------------------------------------------------------
grid <- load_grid()
## Grid correctly loaded for GSA10
```

The grid uses the codes for the countries reported in the following table, derived from the MEDITS manual. In case a country code was not present in the MEDITS protocol the country code was derived from the list of the ISO 3166-1 alpha-3 codes (https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3). Check the country codes used in the TX files and, if the code is different, correct the code reported in the grid (~/scripts/utilities/GRID/GRID_(COUNTRY).csv)

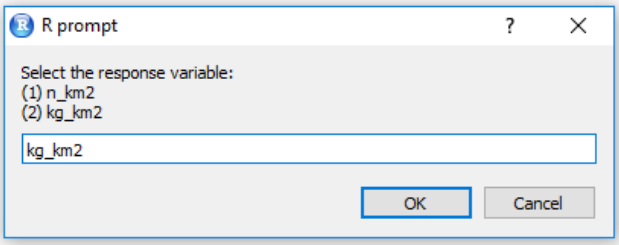
n	CODE	COUNTRY	SOURCE
1	ALB	Albania	MEDITS
2	BGR	Bulgaria	ISO 3166-1
3	CYP	Cyprus	MEDITS
4	DZA	Algeria	ISO 3166-1
5	EGY	Egypt	ISO 3166-1
6	ESP	Spain	MEDITS
7	FRA	France	MEDITS
8	GEO	Georgia	ISO 3166-1
9	GRC	Greece	MEDITS
10	HRV	Croatia	MEDITS
11	ISR	Israel	ISO 3166-1
12	ITA	Italia	MEDITS
13	LBN	Lebanon	ISO 3166-1

n	CODE	COUNTRY	SOURCE
14	LBY	Libya	ISO 3166-1
15	MLT	Malta	MEDITS
16	MON	Montenegro	MEDITS
17	MOR	Morocco	MEDITS
18	PSE	Palestine	ISO 3166-1
19	ROU	Romania	ISO 3166-1
20	RUS	Russia	ISO 3166-1
21	SVN	Slovenia	MEDITS
22	SYR	Syria	ISO 3166-1
23	TUN	Tunisia	ISO 3166-1
24	TUR	Turkey	ISO 3166-1
25	UKR	Ukraine	ISO 3166-1

Running the following code a box will pop-up for the selection of the response variable to be used during the standardization. “kg_km2” is set as default value.

ATTENTION: it is possible to insert only the following values:

- “1” or “n_km2” to select the abundance index
- “2” or “kg_km2” to select the biomass index




```
#---------------------------------------------------------------
# selection of the response variable
#---------------------------------------------------------------
response_variables <- select_response_variable()
## [1] "Select the response variable"
response <- response_variables[[1]]
dependent <- response_variables[[2]]
```

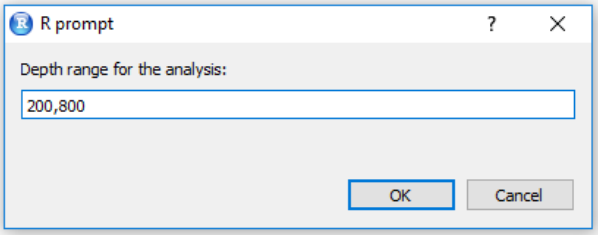
A meta-data base is generated and a summary of the content is outputted. Run the following code:

```
#---------------------------------------------------------------
# creation of metaDB
#---------------------------------------------------------------
data <- create_metaDB()
                 id      COUNTRY        GSA      vessel         year          month
 10ITA2007_PEC811_1 :  1   ITA:500   Min.   :10   PEC:500   Min.   :2007   Min.   :5.000
 10ITA2007_PEC811_2 :  1             1st Qu.:10             1st Qu.:2009   1st Qu.:6.000
 10ITA2007_PEC811_3 :  1             Median :10             Median :2012   Median :7.000
 10ITA2007_PEC811_4 :  1             Mean   :10             Mean   :2012   Mean   :7.022
 10ITA2007_PEC811_46:  1             3rd Qu.:10             3rd Qu.:2014   3rd Qu.:8.000
 10ITA2007_PEC811_5 :  1             Max.   :10             Max.   :2016   Max.   :9.000
 (Other)            :494
      hour            haul          n_haul      duration       GENUS     SPECIES
 Min.   : 4.00   Min.   : 1.0   Min.   :50   Min.   :30.00   -1  :315   -1 :315
 1st Qu.: 7.00   1st Qu.:13.0   1st Qu.:50   1st Qu.:30.00   ARIS:185   FOL:185
 Median :11.00   Median :25.5   Median :50   Median :60.00
 Mean   :10.31   Mean   :25.5   Mean   :50   Mean   :47.42
 3rd Qu.:14.00   3rd Qu.:38.0   3rd Qu.:50   3rd Qu.:60.00
 Max.   :18.00   Max.   :50.0   Max.   :50   Max.   :60.00

       Y               X             depth          stratum         response
 Min.   :38.01   Min.   :13.65   Min.   : 15.0   Min.   :1.000   Min.   : 0.000
 1st Qu.:39.43   1st Qu.:14.29   1st Qu.:111.0   1st Qu.:3.000   1st Qu.: 0.000
 Median :39.97   Median :14.86   Median :331.2   Median :4.000   Median : 0.000
 Mean   :39.93   Mean   :14.95   Mean   :328.9   Mean   :3.538   Mean   : 6.395
 3rd Qu.:40.47   3rd Qu.:15.74   3rd Qu.:581.6   3rd Qu.:5.000   3rd Qu.:10.339
 Max.   :41.14   Max.   :16.04   Max.   :672.0   Max.   :5.000   Max.   :77.513
```

Select the bathymetric range for the analysis from the box that pops up running the following code. The default values are "10,100" m but it is possible to select any interval composed by the contiguous strata reported in the file:
"~/scripts/utilities/strata.csv".

ATTENTION: check that the strata.csv file in the utilities folder is compiled with the strata used in the sampling. Compiling the box, separate the values only with comma.

```
#-------------------------------------------------------------
# depth range selection
#-------------------------------------------------------------
source(paste(wd,"/standardization/depth_range.r", sep=""))
## [1] "Select the depth range for the analysis"
```



The routine allows performing an exploration of data applying the Variance Inflation Factors (VIF) to test the collinearity among the variables. Moreover, the correlation matrix is computed and outputted. Results of the explorative analysis are saved in the following files:

- ARISFOL_GSA10_(biomass)___200-800__VIF.txt
- ARISFOL_GSA10_(biomass)___200-800__Correlation_Matrix.txt

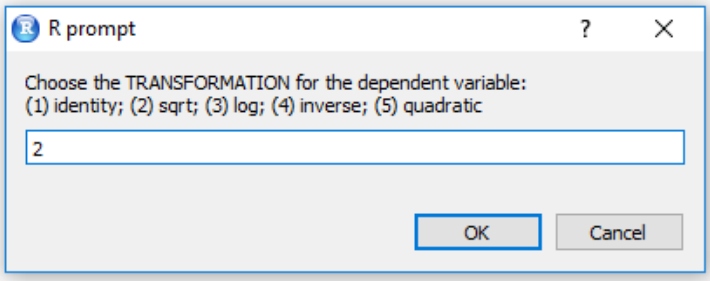
```
#-------------------------------------------------------------
# explorative analysis of data
#-------------------------------------------------------------
source(paste(wd,"/standardization/explorative_analysis.r", sep=""))
var   vif
year  1.02466058389794
month 1.02172768246888
hour  1.0100048154028
Y     22.2289189689707
X     19.7788804801777
depth 1.83633907547682

removed:  Y 22.22892

year month  hour     Y     X depth
year   1.00  0.14 -0.05  0.00  0.00  0.03
month  0.14  1.00  0.02  0.04 -0.03 -0.03
hour  -0.05  0.02  1.00  0.02 -0.01  0.02
Y      0.00  0.04  0.02  1.00 -0.96 -0.39
X      0.00 -0.03 -0.01 -0.96  1.00  0.22
depth  0.03 -0.03  0.02 -0.39  0.22  1.00

n= 290


P
year   month  hour   Y      X      depth
year         0.0187 0.4064 0.9441 0.9779 0.6613
month 0.0187        0.7775 0.5281 0.5803 0.6396
hour  0.4064 0.7775        0.6826 0.8229 0.7052
Y     0.9441 0.5281 0.6826        0.0000 0.0000
X     0.9779 0.5803 0.8229 0.0000        0.0001
```

```
depth 0.6613 0.6396 0.7052 0.0000 0.0001
```

Different transformations could be applied to the response variable to improve the model fitting. Running the following command a box will pop up for the selection of the transformation to be applied. Choose one of the four possibilities using either numeric or string values (e.g. select "2" or "sqrt" to follow the example reported in the present tutorial):

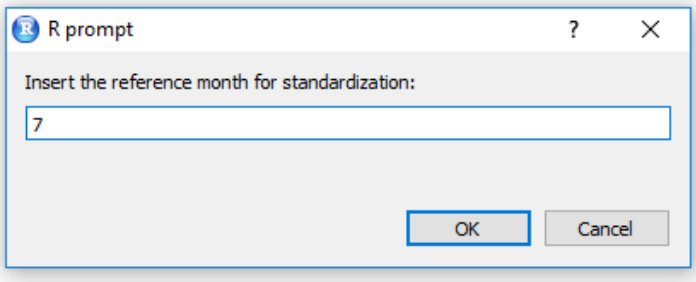
1. identity: no transformation
2. sqrt: $\sqrt{response}$
3. log: $log_e(response + 1)$
4. inverse: 1/response
5. quadratic: $(response^2)$

```
#---------------------------------------------------------------
# transformation of response variable
#---------------------------------------------------------------
source(paste(wd,"/standardization/transformation.r", sep=""))
'sqrt' transformation applied to the response variable.
```



Whether the month variable is used in the final model, a reference month should be selected for the standardization. In the pop up window insert only the numerical value referred to the chosen month. The default value is 7 for the month of July.

```
#---------------------------------------------------------------
# choise of the refernce month for standardization
#---------------------------------------------------------------
month_st <- month_resp()
```



At this point it is possible to start the selection of the best model using the preferred method such as, for example, the forward or backward stepwise selection.
The routine reports only the first step of the forward inclusion procedure. It consists in fitting the basic models using only one explanatory variable at a time. At each step is necessary to inspect the summary of the basic models (*summary(mod1)*, *summary(mod2)*, and so on). Choose the model with the higher explained deviance, with significant explanatory variable and the lowest Generalized Cross-Validation (GCV) value. In case of models using distributions not belonging to the exponential family (such as the Tweedie distribution, generally used in case of zero-inflated datasets) the REML method is adopted. In these cases the GCV method is no more effective and it is suggested to compare the models selecting the ones with the lowest AIC value (Akaike Information Criterion: e.g. "*AIC(mod1)*"). At each step, add to the selected model another variable at the time. The computed splines on the explanatory variables should be inspected in order to retain only the ones making sense from an ecological and biological point of view. In case the degree of the spline seems to be too high, producing too many oscillations difficult to be biologically explained, the degree of the spline can be constrained, fixing in the corresponding spline the k value to be used (e.g. *s(depth, k=7)*).

In order to select the variables to be used in the modelling process, it is possible to inspect the composition of the data frame running the following command:

```
#------------------------------------------------------------
#------------------------------------------------------------
# Stepwise selection of GAM models
#------------------------------------------------------------
#------------------------------------------------------------
variables <- data.frame(variables = colnames(data)); variables
##    variables
## 1         id
## 2    COUNTRY
## 3        GSA
## 4     vessel
## 5       year
## 6      month
## 7       hour
## 8       haul
## 9     n_haul
## 10  duration
## 11     GENUS
## 12   SPECIES
## 13         X
## 14         Y
## 15     depth
## 16   stratum
## 17  response
```

Run the following command to estimate the basic models:

```
mod1 <- gam(response ~ s(X)+0, family= gaussian (link = identity),data=data, se
lect=T, gamma = 1.4)
mod2 <- gam(response ~ s(Y)+0, family= gaussian (link = identity),data=data, se
lect=T, gamma = 1.4)
mod3 <- gam(response ~ s(depth)+0, family= gaussian (link = identity),data=data
, select=T, gamma = 1.4)
```

```
mod4 <- gam(response ~ s(year)+0, family= gaussian (link = identity),data=data,
 select=T, gamma = 1.4)
mod5 <- gam(response ~ factor(month)+0, family= gaussian (link = identity),data
=data, select=T, gamma = 1.4)
mod6 <- gam(response ~ s(hour)+0, family= gaussian (link = identity),data=data,
 select=T, gamma = 1.4)
```

and then inspect the model summary:

```
# GCV and explained deviance can be inspected from the summary of the models
summary(mod1)
##
## Family: gaussian
## Link function: identity
##
## Formula:
## response ~ s(X) + 0
##
## Approximate significance of smooth terms:
##        edf Ref.df     F p-value
## s(X) 1.448      9 0.387   0.085 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0202   Deviance explained = 1.68%
## GCV = 1.4609e+06  Scale est. = 1.4478e+06  n = 290
… … …
# once that the best basic model was detected, the variables should be included
 one at the time in the same way
#.....
# repeat until the best model was detected
```

Once that the variables for the best model are selected, before running the final model it is necessary to declare the family and the link function to be used:

```
#-------------------------------------------------------------
#-------------------------------------------------------------

tranf_type
[1] "sqrt"
family <- "gaussian"  # "quasipoisson"     "tw"
linkfun <- "identity" # "log"   "inverse
formula_best <- "response ~ s(Y) + s(depth) + factor(year)"
```

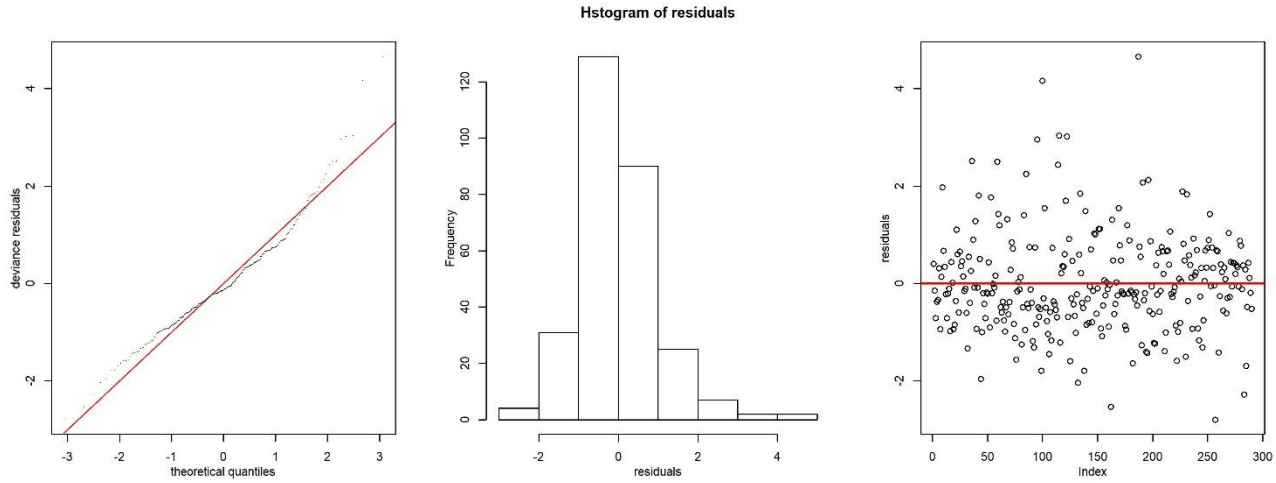
Run the following codes to estimate the final model and inspect the summary:

```
mod <- mod_estimation(formula_best,family, linkfun)
summary(mod)

Family: gaussian
Link function: identity

Formula:
response ~ s(X,Y) + s(depth) + factor(year)
```

```
Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.1393736  0.1947551  10.985  < 2e-16 ***
factor(year)2008  0.5607225  0.2752010   2.038 0.042581 *
factor(year)2009  0.8824819  0.2753227   3.205 0.001513 **
factor(year)2010  0.9236641  0.2751737   3.357 0.000903 ***
factor(year)2011  0.7976124  0.2754881   2.895 0.004100 **
factor(year)2012  0.3808768  0.2760913   1.380 0.168879
factor(year)2013  0.5846111  0.2755804   2.121 0.034807 *
factor(year)2014 -0.3380300  0.2752179  -1.228 0.220438
factor(year)2015 -0.2820860  0.2750756  -1.025 0.306058
factor(year)2016 -0.0001461  0.2752149  -0.001 0.999577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
           edf Ref.df      F  p-value
s(X,Y)   7.508     27  0.807 0.000966 ***
s(depth) 4.164      9 64.074  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.774   Deviance explained =   79%
GCV = 1.2628  Scale est. = 1.0941    n = 290
```

To continue with the inspection of residuals run the following code:

```
#-------------------------------------------------------------
# Inspection of residuals of the final model
#-------------------------------------------------------------
residuals_inspection()
```

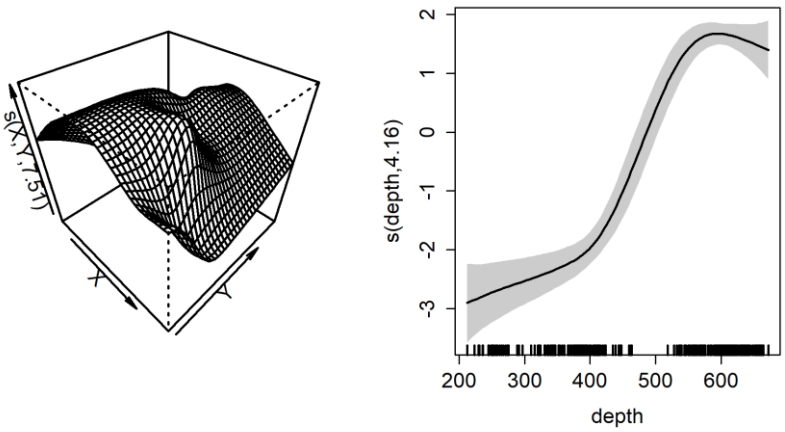


The plot is saved in the *output_GAM* folder with the following name:

- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity__200-800_Residuals.jpg

The plot of the splines of the models is produced running the following command:

```
#-------------------------------------------------------------
# plot of slines
#-------------------------------------------------------------
plot_splines()
```

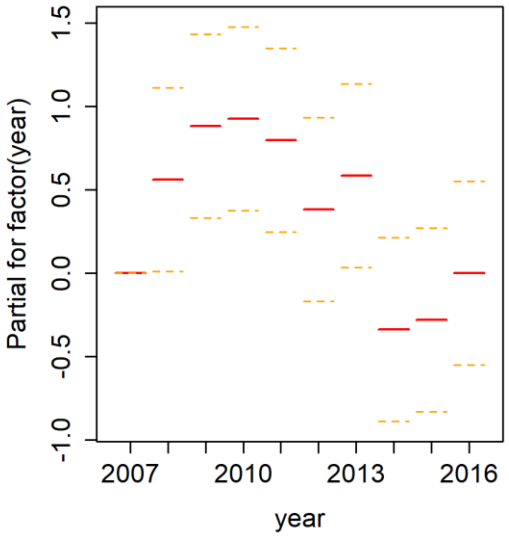


The plot is saved in the *output_GAM* folder with the following name:

- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity__200-800_splines.jpg

In case one or more variables were used as factors, is possible to produce the plot running the following command:

```
#-------------------------------------------------------------
# plot of factors
#-------------------------------------------------------------
plot_factors()
```

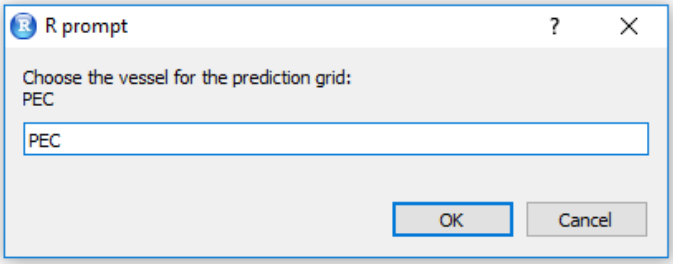


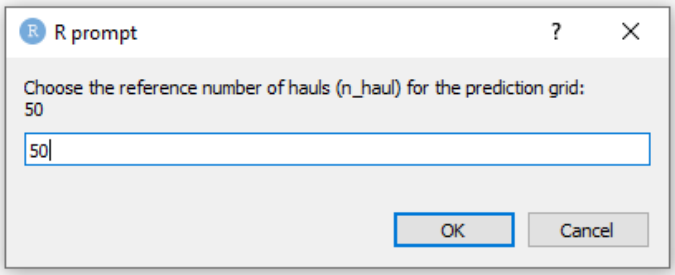
The plot is saved in the *output_GAM* folder with the following name:

- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity__200-800_factors.jpg

The routine predicts the annual indices as the mean index by depth stratum. The annual index for the study area is computed as the mean of these values by stratum, weighted by the stratum weight (surfaces), according to Souplet (1996). In case a transformation was applied to the data before modelling or in case of using a link function different from identity, the routine re-transform the indices estimated by the model. The time series of the observed indices, previously estimated by the *BioIndex* routine are plotted together with predicted indices.

Running the following code, in case "vessel" and "n_haul" variables were used in the final model, the following boxes will pop up to allow the selection of the reference vessel and number of hauls in the annual survey to be used during the standardization:

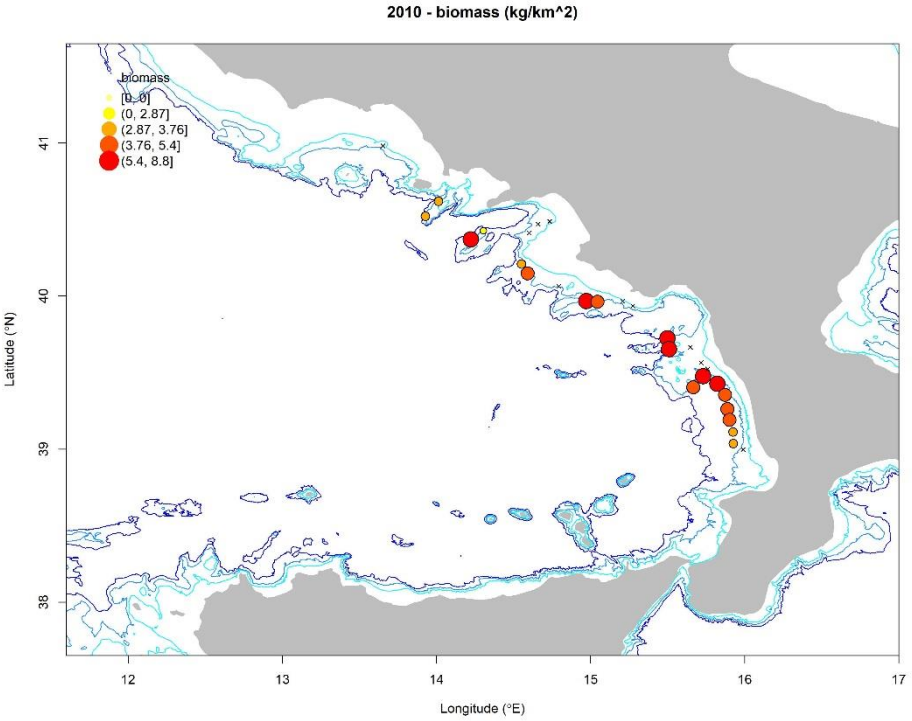


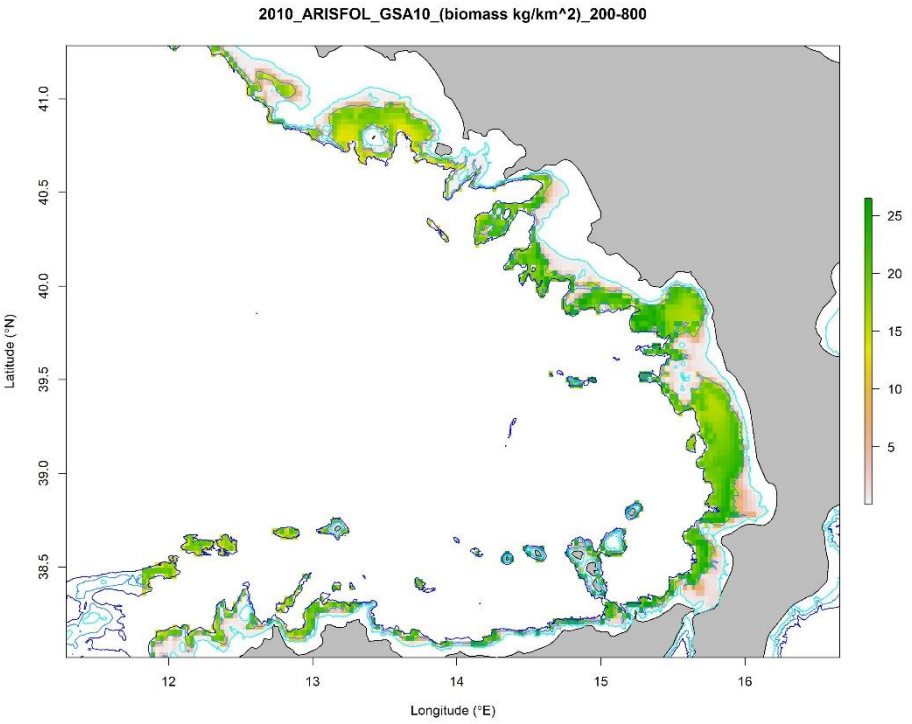



```
#------------------------------------------------------------
# Prediction of the standardized indices
#------------------------------------------------------------
source(paste(wd,"/standardization/Index_Prediction_month.R", sep=""), encoding
= 'UTF-8')
```

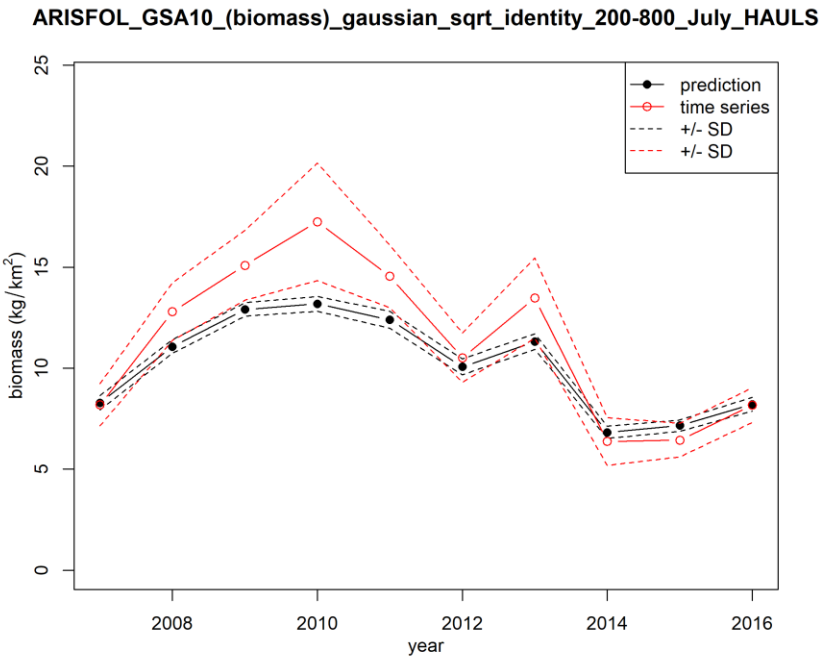
The routine also provides the possibility to plot both the annual maps of the final model's prediction and of the bubble plots of the observed data ("map" folder). The raster file of each annual map is also saved in the "raster" folder. The maps could be useful to assess the quality of the prediction returned by the final model. The user have to answer to the following question:

```
Would you like to save the prediction maps for each year?

1: Yes
2: No

Selection:
```





The bathymetrical lines indicates respectively: 200m of depth (cyan); 500m of depth (light blue); 800m of depth (dark blue).




```
Family: gaussian
Link function: identity

Formula:
response ~ s(X,Y) + s(depth) + factor(year)

Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.1393736  0.1947551  10.985  < 2e-16 ***
factor(year)2008  0.5607225  0.2752010   2.038 0.042581 *
factor(year)2009  0.8824819  0.2753227   3.205 0.001513 **
factor(year)2010  0.9236641  0.2751737   3.357 0.000903 ***
factor(year)2011  0.7976124  0.2754881   2.895 0.004100 **
factor(year)2012  0.3808768  0.2760913   1.380 0.168879
factor(year)2013  0.5846111  0.2755804   2.121 0.034807 *
factor(year)2014 -0.3380300  0.2752179  -1.228 0.220438
factor(year)2015 -0.2820860  0.2750756  -1.025 0.306058
factor(year)2016 -0.0001461  0.2752149  -0.001 0.999577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
           edf Ref.df      F  p-value
s(X,Y)   7.508     27  0.807 0.000966 ***
s(depth) 4.164      9 64.074  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.774   Deviance explained =   79%
GCV = 1.2628  Scale est. = 1.0941    n = 290
```

The following files are saved in the output_GAM folder

- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity_200-800_July_GRID_prediction.tiff
- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity_200-800_July_GRID_prediction.csv
- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity_200-800_July_HAULS_prediction.tiff
- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity_200-800_July_HAULS_prediction.csv
- ARISFOL_GSA10_(biomass)_gaussian_sqrt_identity__200-800__summary.txt

References

Souplet, A. (1996). Définition des estimateurs. In: Campagne internationale de chalutage démersal en Méditerranée (Medits 95). Vol. III. Indices de biomasse et distributions en tailles. Bertrand J. Coordonnateur général. Etude 94/047 IFREMER/CE, 94/011 IEO/CE, 94/057 SIBM/CE, 94/051 NCMR/CE