SampleOptim Documentation

Patrícia Gonçalves (IPMA), email: patricia@ipma.pt

1. Introduction

The SampleOptim R-toolbox (SampleOptimRDB) aims to provide quality indicators estimations to support user's decision to determine the optimal number sample size for age-length keys and maturity ogives.

The simulation process works at sample level.

The dataset used represents the "whole" population, and the simulations are based on subsampling without replacement. Although, SampleOptim allow the user to define the type of subsampling with or without replacement.

The subsamples obtained from the simulations intend to allow the comparison of the ALKs and of the maturity ogive parameters based on a reduction on the number of individuals sampled by length class.

The setup for simulations allows the user to perform those type of scenarios taking into account a length stratification together with other possible stratifications:

- Temporal stratification (annual, semester and quarter);
- Sex stratification, by defining the sexratio on the subsamples (i.e. proportion of females and males by length class);
- Port stratification (harbour of the samples provenience), options are using a randomly sampling by port, or define a uniform sample distribution by port.

2. R-packages required

The R-packages required are: FSA, FSAdata, nlstools, reshape, ggplot2, ggthemes, cvTools, dplyr, robustbase, MASS, psyphy, boot and RCurl.

3.Input data

3.1.1 Dataset format

The dataset required to run the simulations should contain the following mandatory columns:

- ID_BIO_FISH;
- date;
- month;
- year;
- COD_FAO;

- Port;
- Length_class;
- Weight;
- Sex;
- Maturity_stage;
- Age

Each line of this dataset corresponds to an individual.

A function which makes the conversion of the CS-CA (RDB exchange format) into this input dataset are available in the script named as: “A.Data conversion and preparation.R”

3.1.2 Description of the variables on the dataset (described in Section 3.1.1)

$ID_BIO_FISH : int (unique and different numeric value for each line)

$date : Factor

$month : int (month of sampling; numeric value from 1 to 12)

$year : int (year of sampling, four numbers format)

$COD_FAO : int (three letter FAO code for the species)

$Port : Factor (Port/harbour name; or Region/Area of sampling)

$Length_class : int (length class measurement, units: mm, cm, m)

$Weight : num (units: Kg, g)

$Sex : Factor (F - females; M - males; I - undetermined sex)

$Maturity_stage: int or character (according to the species maturity stage key)

$Age : int

3.2. Initial values

Initial values necessary to be provide in an .csv (file name: input_params.csv) file to define the setup of the simulations scenarios:

Names.of.variables	Mandatory	Variable.options	Default	Definition
species	y			CODE_FAO
AREA	y			
VARIABLE	y	age; all	all (age and maturity)	age - only statistical analysis for age; all - statistical analysis age and maturity

Names.of.variables	Mandatory	Variable.options	Default	Definition
PORT*	n	TRUE; FALSE	FALSE	Uses Port stratification for subsampling (TRUE); Do not consider Port stratification for subsampling (FALSE)
distUniPorto	n	TRUE; FALSE	FALSE	Uniform distribution of subsamples by Port (TRUE); Randomly distribution of subsamples by Port (FALSE)
TIME_STRATA	y	A; S; T		A - year; S - semester; T - quarter
SEX_RATIO	y	0;1; 0<numeric<1; FALSE		0 - only males; 1 - only females; 0<numeric<1 sexratio proportion; FALSE - not considers sexratio
MIN_LC	y			minimum length class
MAX_LC	y			maximum length class
interval_LC	y	numeric		length class step
MIN_age	y			minimum age
MAX_age	y			maximum age
MIN_OTOL.Read	y			minimum number of individuals by length class
MAX_OTOL.Read	y			maximum number of individuals by length class
interval_OTOL.Read	y			interval number of individuals by length class in the simulation setup
Linf	y			von Bertallanfy growth model parameter - Linf. Used as a starting value to adjust VBGM.
K	y			von Bertallanfy growth model parameter - k. Used as a starting value to adjust VBGM.
t0	y			von Bertallanfy growth model parameter - t0. Used as a starting value to adjust VBGM.
year_start	y			first year data subset to run simulations

Names.of.variables	Mandatory	Variable.options	Default	Definition
year_end	y			last year data subset to run simulations
stage_mature	y	numeric		define the maturity stages that correspond to mature stages (to allow to determine the proportion of immatures and matures)
n	y	numeric		define the number of simulations (bootstrap runs)

* in cases here the PORT is not important to be considered for stratification, the variable could be replaced for the FLEET. If the FLEET is considered, the option “distUniPorto” will be used to: Uniform distribution of subsamples by fleet (when settled as: TRUE); Randomly distribution of subsamples by fleet (when settled as: FALSE).

4. Scripts files

4.1 “00.Data conversion and preparation.R”

Converts the CSCA.csv RDB data file into the required dataset format (described in Sections 3.1.1 and 3.1.2).

Since the main aim of the simulations scenarios are deciding on the “optimal” number of individuals by length class for ALKs, the dataset should only include samples with individuals with age readings.

4.2 Script file: “1_Data_exploratory_analysis_RDB.R”

Exploratory analysis of the selected dataset to simulations are performed. The main analysis are:

- Table with the number of samples by Port, year and month;
- Table length classes of the samples by Port, year and month;
- Table age of the samples by Port, year and month;
- Plot with length distribution samples by Port by year and month;
- Plot age distribution samples by Port by year and month;
- Plot length distribution by year;
- Plot age distribution by year;

In the future: will be added more plots and table options to explore the sampling data.

4.3 Script file: “2_Simulations_RDB.R”

Simulation scenarios were settled up here, which means the number of individuals by length class to be evaluated were defined as the base for the simulation procedure. The simulation stratification options for subsampling based on: temporal variables (year, semester and quarter), ports and sexes (sexratio). For each scenario the simulation algorithm number of runs are defined by the user and also the option of sampling with or without replacement. The simulation data operational model is: “sample_selection_function_RDB.R” The data subsets from each scenario are saved in .csv files.

The data from each data subsets and from each simulation run, are then analysed and the following parameters determined:

- L75, L50 and L25 (the length at which 75%, 50% and 25% of the individuals are mature) and the respective confidence intervals;
- standard deviation of length at age for the original dataset with the data subsets from the simulations;
- comparison standard deviation of length at age from distributions of the original dataset with the data subsets from the simulations;
- von Bertallanfy growth model (VBGM) parameters (Linf, K, t0) of each data subset from the simulations;
- root trimmed mean squared percent error (RTMSPE) obtained via (repeated) K-fold cross-validation;
- mean squared percent error (MSPE) obtained via (repeated) K-fold cross-validation;
- mean absolute percent error (MAPE) obtained via (repeated) K-fold cross-validation.

Note: for the determination of the maturity ogive parameters (L75, L50 and L25) the R-script has an option to constraint the data subset to the individuals collected only during the main spawning period of the species.

4.4 Script file: “3_Simulations_results_data_analysis_RDB.R”

The results from the different scenarios and simulation runs are compiled here at this stage. Most of the data compilation at this stage is based on the production of the following figures with:

- comparison of the VBGM parameters between simulation runs and scenarios;
- comparison of the mean length-at-age, standard deviation length-at-age and the coefficient of variation length-at-age from the original dataset with the data subsets from the simulations runs and scenarios;
- comparison of the mean age-at-length, standard deviation age-at-length and the coefficient of variation age-at-length from the original dataset with the data subsets from the simulations runs and scenarios;

- comparison of the different stats (RTMSPE, MSPE and MAPE), used as quality indicators (Qis), of the original dataset with the data subsets from the simulations runs and scenarios.

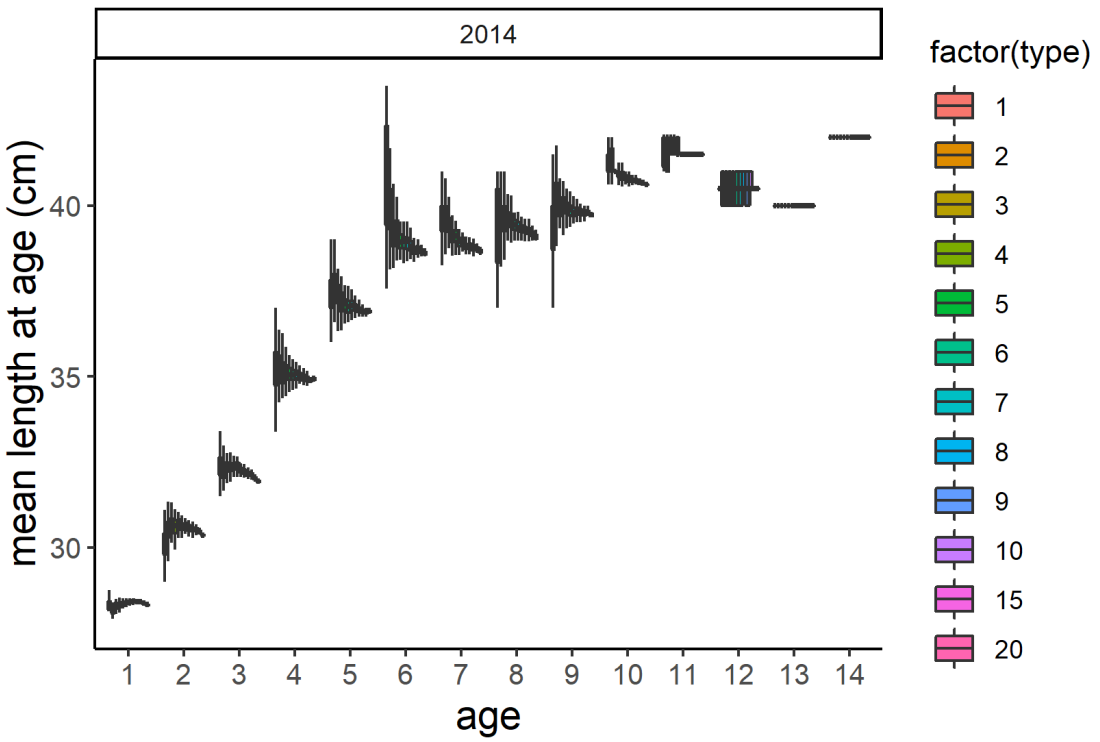
The decision on the optimum number of individuals to be sampled by length class is based on visual interpretation of those figures.

4.5 Outputs

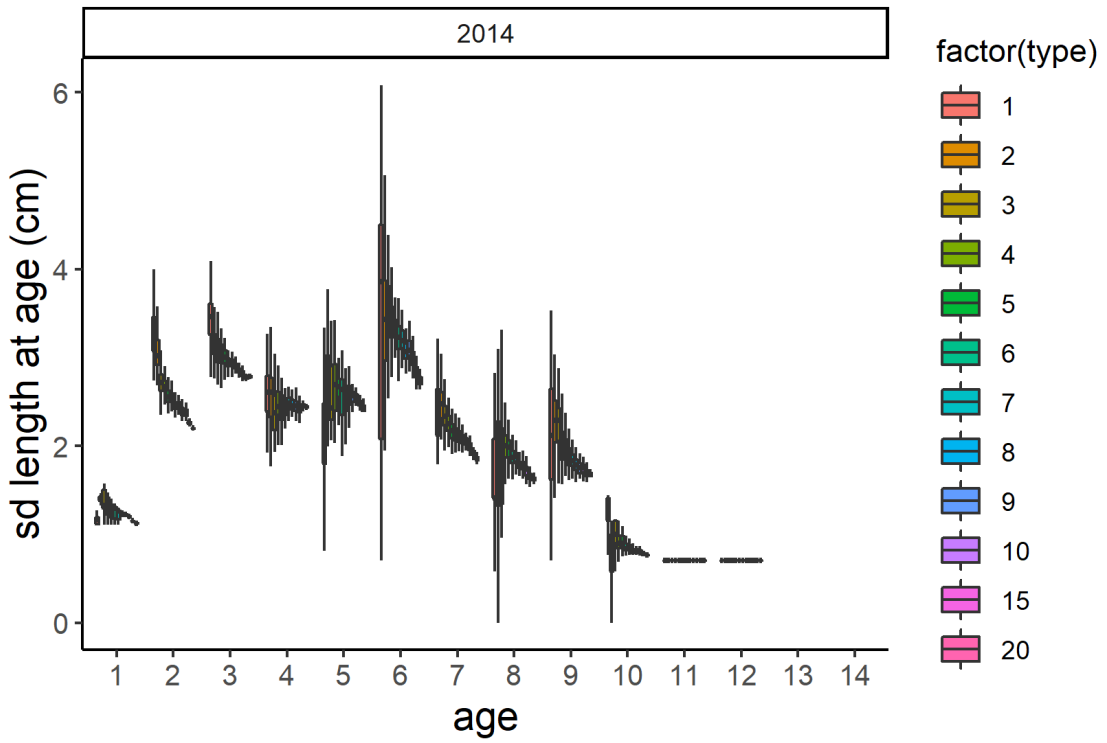
An example of some of the main outputs visual (section 4.5.1) and summary data tables (section 4.5.2) produced by SampleOptim are presented in the next sections.

4.5.1 Visual

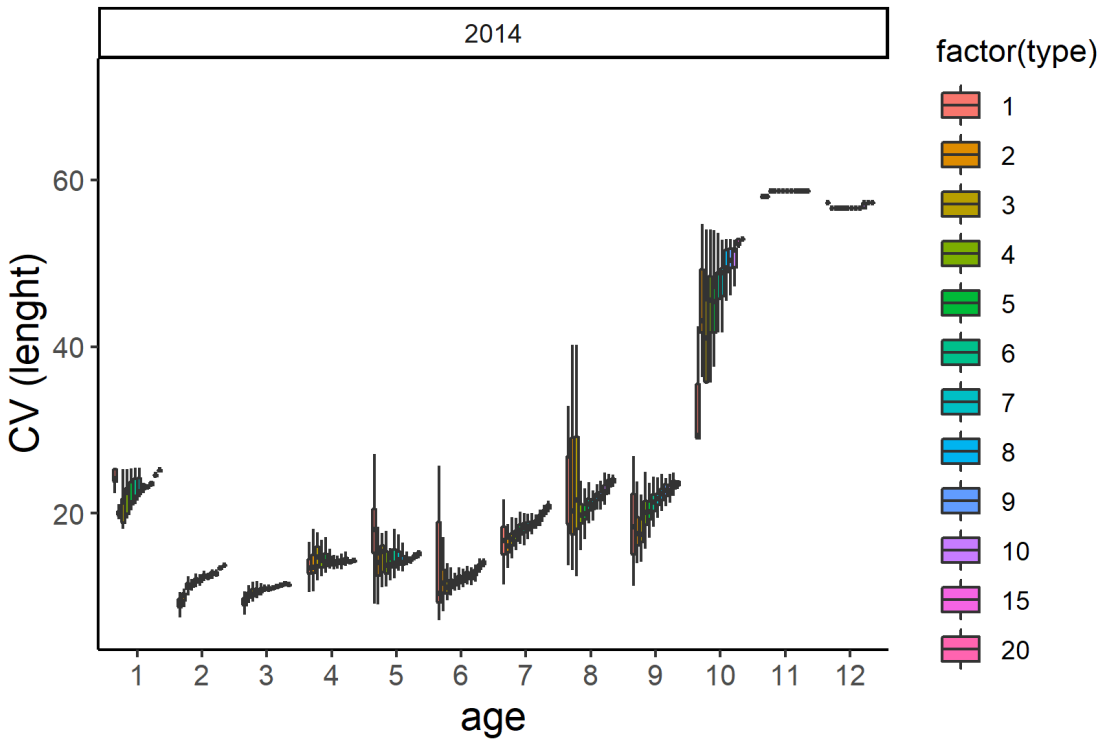
Example of a boxplot with the mean length-at-age by simulation condition. The factor(type) indicates the number of individuals sampled by length class.



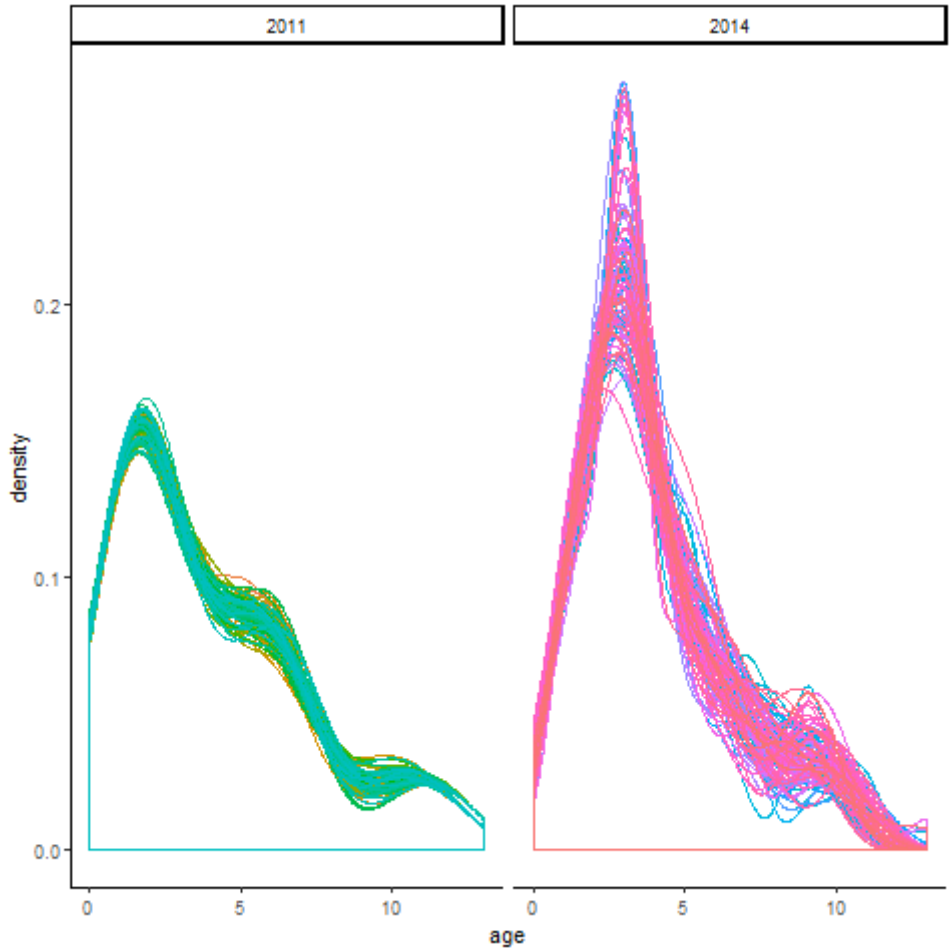
Example of a boxplot with the standard deviation length-at-age by simulation condition. The factor(type) indicates the number of individuals sampled by length class.



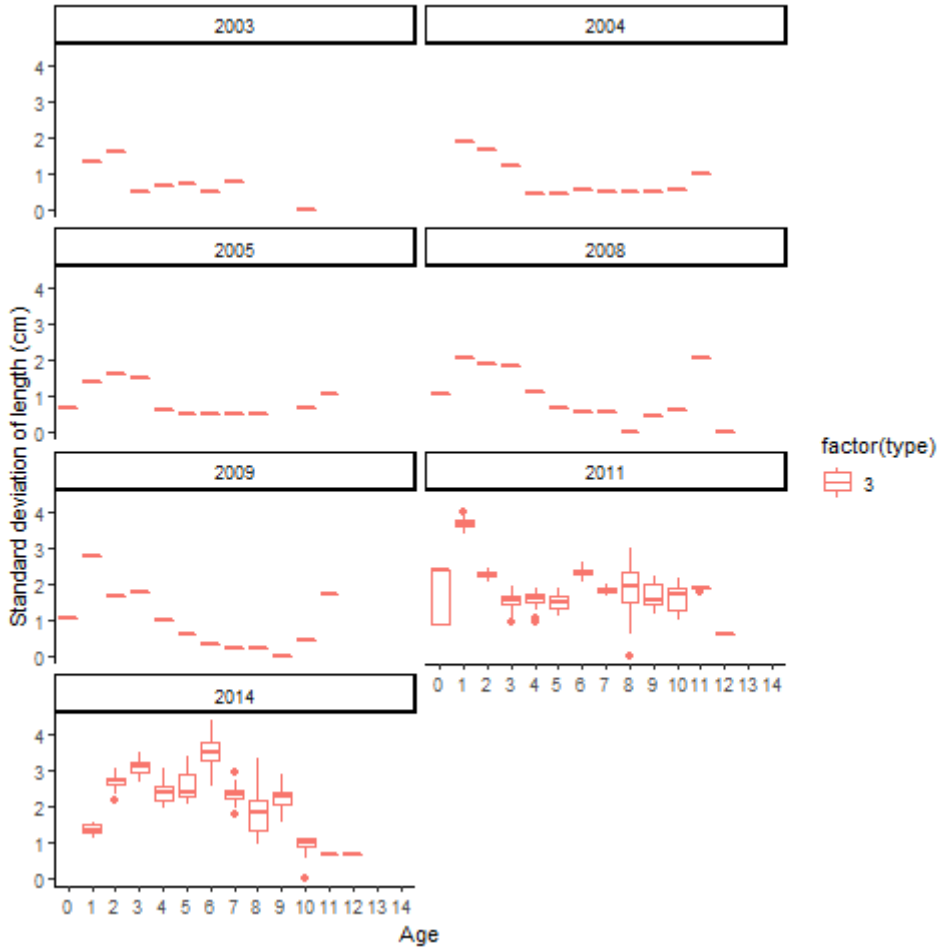
Example of a boxplot with the coefficient of variation length-at-age by simulation condition. The factor(type) indicates the number of individuals sampled by length class.



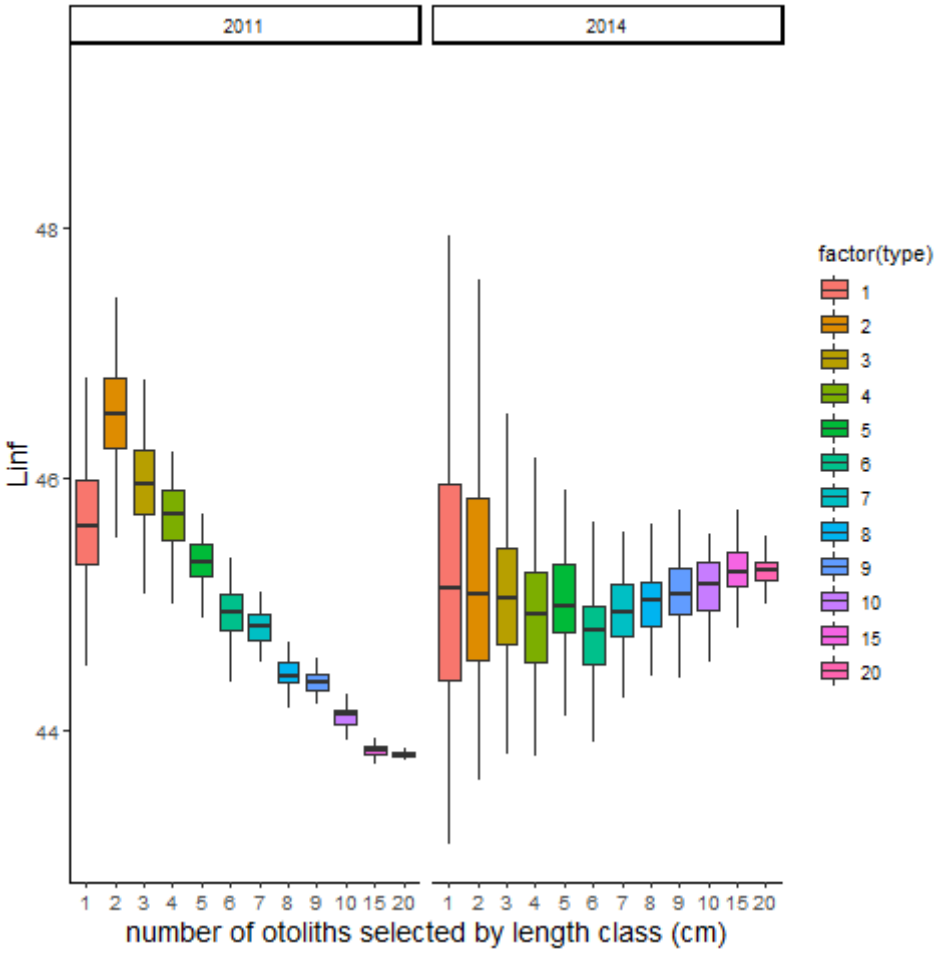
Example of a density plot of ages, based on simulation results testing the hypothesis of sampling 1 individual by length class and quarter for the dataset from 2011 and 2014. Each line corresponds to a simulation run from 1 to 100.



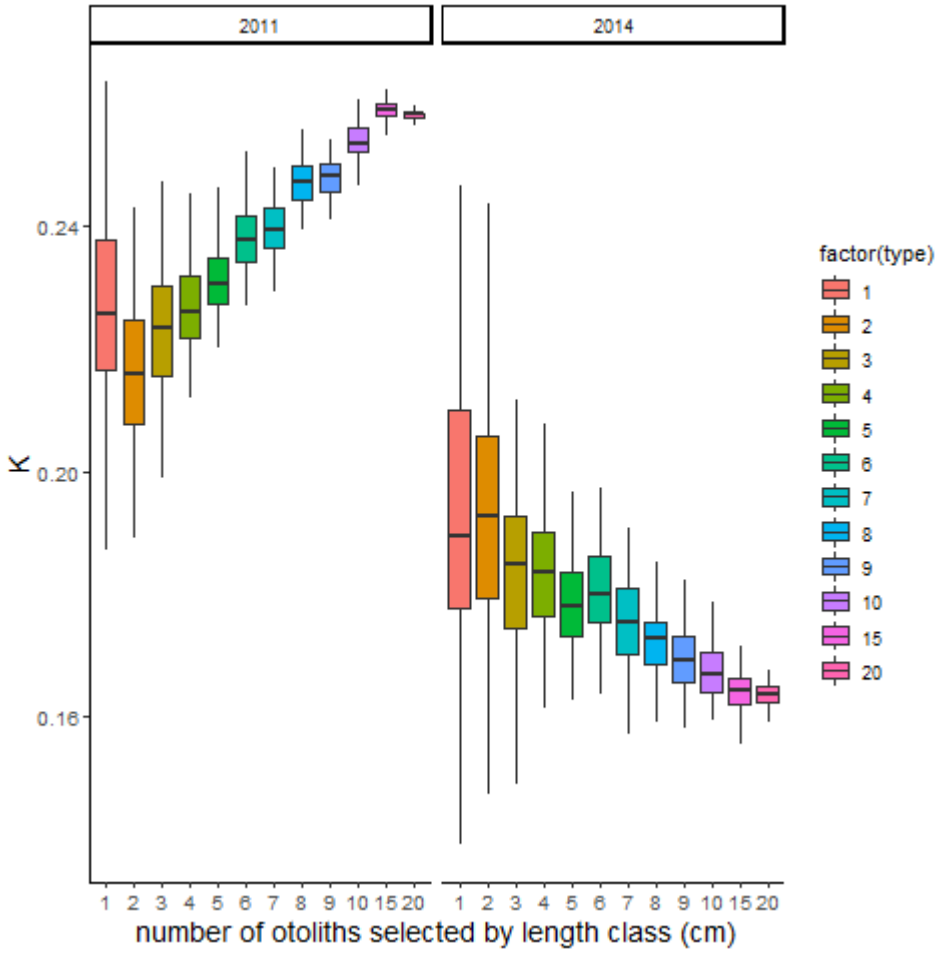
Example of a boxplot of the standard deviation of length at each age class, considering the data resulting from the simulation runs (n=100) based on the condition of sampling 3 individuals by length class (factor(type)).



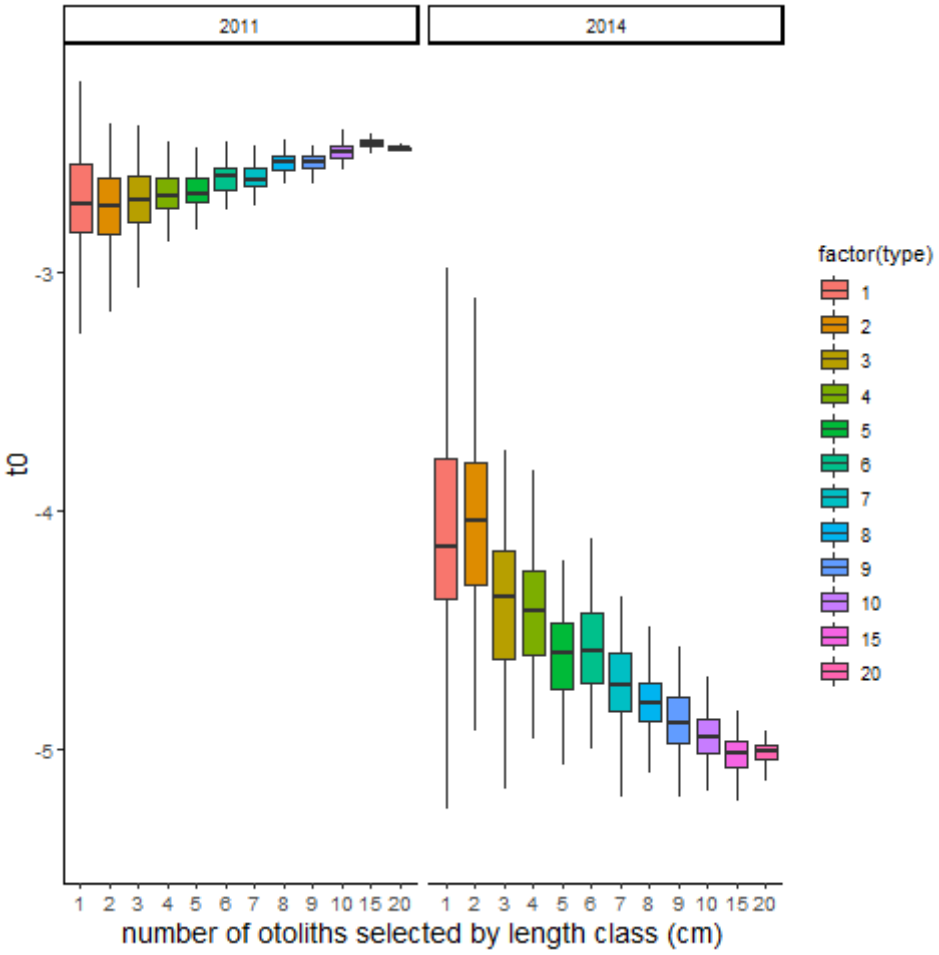
Example of a boxplot with the values of the von Bertalanffy growth model parameter Linf, by simulation condition. The factor(type) indicates the number of individuals/otoliths sampled by length class.



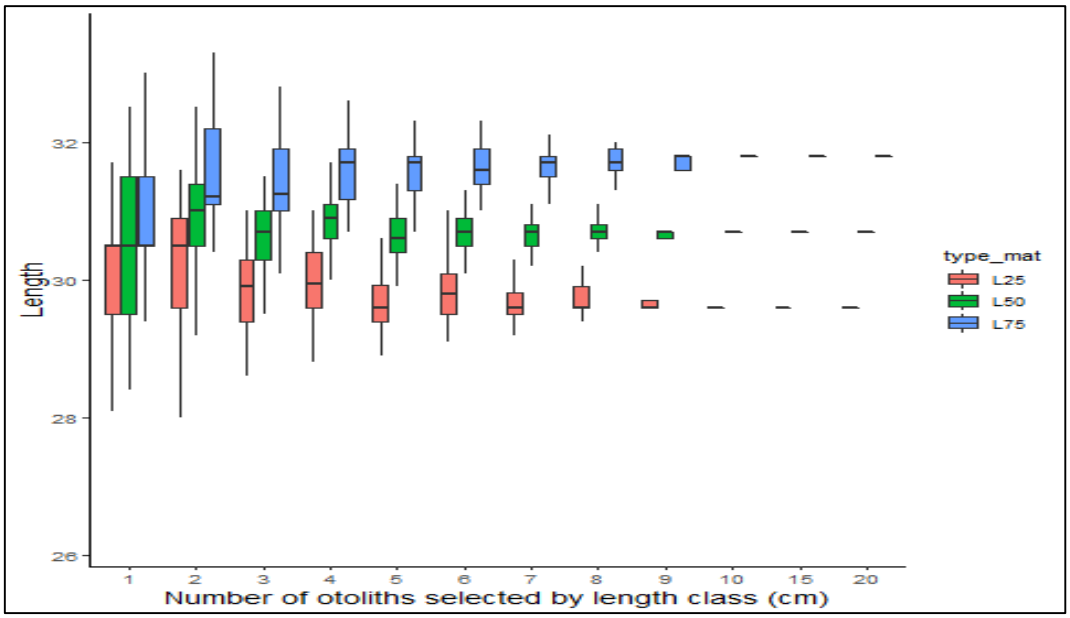
Example of a boxplot with the values of the von Bertalanffy growth model parameter K, by simulation condition. The factor(type) indicates the number of individuals/otoliths sampled by length class.



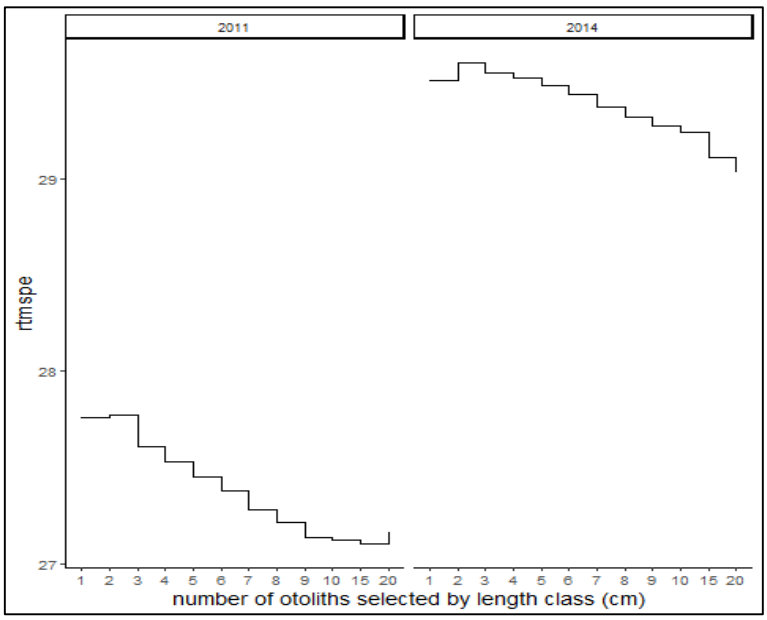
Example of a boxplot with the values of the von Bertalanffy growth model parameter t0, by simulation condition. The factor(type) indicates the number of individuals/otoliths sampled by length class.



Example of a boxplot with the values of the adjustment of the maturity ogive (L25, L50, L75), determined by simulation condition according to the number of individuals/otoliths sampled by length class.



Example of a plot with the root mean squared prediction error (rtmspe) by the number of individuals/otoliths sampled selected by length class.



4.5.2 Summary tables

Example of an output table with the values of Linf, k and t0 determined by simulation run (from 1 to 100; ID_sim) and according to the simulation condition, in this particular case 8 individuals by length class (type).

	A	B	C	D	E	F	G	H
1	Linf,"K","t0","year","ID_sim","type"							
2	1,44.3189365651322,0.252229094321758,-2.48287194226823,2011,1,8							
3	487,44.3701837917741,0.247956421627995,-2.55063787330413,2011,2,8							
4	973,44.3771125541627,0.251110329890265,-2.50611052174692,2011,3,8							
5	1459,44.5852060556506,0.2452254313856,-2.53636275591729,2011,4,8							
6	1945,44.4092668647964,0.247295576927735,-2.54834608959658,2011,5,8							
7	2431,44.5336159496939,0.245788008664458,-2.55835206691808,2011,6,8							
8	2917,44.3237873073551,0.251786575455623,-2.50108016954353,2011,7,8							
9	3403,44.5358658426911,0.243438018711457,-2.5877985051835,2011,8,8							
10	3889,44.3368573479964,0.249776084813108,-2.5218253301027,2011,9,8							
11	4375,44.4098494142301,0.248442647419666,-2.52069434026527,2011,10,8							
12	4861,44.3916930274929,0.248351938879542,-2.54575592947693,2011,11,8							
13	5347,44.359485222405,0.250471475937453,-2.50770802850255,2011,12,8							
14	5833,44.3949358134074,0.247181575023238,-2.54555669125954,2011,13,8							
15	6319,44.4433509165809,0.247929057361271,-2.52662413085017,2011,14,8							
16	6805,44.6273671431361,0.24113290578009,-2.621083592505,2011,15,8							
17	7291,44.3807487953177,0.247597850363377,-2.53706471459613,2011,16,8							

Example of an output table with the values from the adjusting of the maturity ogive (L25, L50, L75) determined by simulation run (from 1 to 100; ID_sim) and according to the simulation condition, in this particular case 15 individuals by length class (type).

	A	B	C	D	E
1	year,"L25","L50","L75","ID_sim","type"				
2	1,2011,12.4,14.5,16.6,1,15				
3	2,2011,12.4,14.5,16.6,2,15				
4	3,2011,12.4,14.5,16.6,3,15				
5	4,2011,12.4,14.5,16.6,4,15				
6	5,2011,12.4,14.5,16.6,5,15				
7	6,2011,12.4,14.5,16.6,6,15				
8	7,2011,12.4,14.5,16.6,7,15				
9	8,2011,12.4,14.5,16.6,8,15				
10	9,2011,12.4,14.5,16.6,9,15				
11	10,2011,12.4,14.5,16.6,10,15				
12	11,2011,12.4,14.5,16.6,11,15				
13	12,2011,12.4,14.5,16.6,12,15				
14	13,2011,12.4,14.5,16.6,13,15				
15	14,2011,12.4,14.5,16.6,14,15				
16	15,2011,12.4,14.5,16.6,15,15				
17	16,2011,12.4,14.5,16.6,16,15				

5. Case studies examples

5.1.1 Case study 1

Species scientific name: *Scomber scombrus*

Species common name: mackerel

Area: ICES 27.9.a

The results from case study 1 have been showed on the presentation made at the first day of ICES WKBIOPTIM3, named as: “SampleOptim R-tool to optimize fish sampling for biological parameters”.

5.1.2 Dataset description

The input/initial values table for the case-study 1 are present here:

Names.of.variables	Mandatory	Variable.values	Definition
species	y	MAC	CODE_FAO
AREA	y	27.9.a	
VARIABLE	y	all	age - only statistical analysis for age; all - statistical analysis age and maturity
PORT	n	TRUE	Uses Port stratification for subsampling (TRUE); Do not consider Port stratification for subsampling (FALSE)
distUniPorto	n	TRUE	Uniform distribution of subsamples by Port (TRUE); Randomly distribution of subsamples by Port (FALSE)
TIME_STRATA	y	T	A - year; S - semester; T - quarter
SEX_RATIO	y	0.5	0 - only males; 1 - only females; 0<numeric<1 sexratio proportion; FALSE - not considers sexratio
MIN_LC	y	13	minimum length class
MAX_LC	y	49	maximum length class
interval_LC	y	1	length class step
MIN_age	y	0	minimum age
MAX_age	y	10	maximum age
MIN_OTOL.Read	y	1	minimum number of individuals by length class

Names.of.variables	Mandatory	Variable.values	Definition
MAX_OTOL.Read	y	20	maximum number of individuals by length class
interval_OTOL.Read	y	1	interval number of individuals by length class
Linf	y	50	von Bertallanfy growth model parameter - Linf. Used as a starting value to adjust VBGM.
K	y	0.1	von Bertallanfy growth model parameter - k. Used as a starting value to adjust VBGM.
t0	y	-3	von Bertallanfy growth model parameter - t0. Used as a starting value to adjust VBGM.
year_start	y	2011	first year data subset to run simulations
year_end	y	2014	last year data subset to run simulations
stage_mature	y	>2	define the maturity stages that correspond to mature stages (to allow to determine the proportion of immatures and matures)
n	y	100	define the number of simulations (bootstrap runs)

5.2 Case study 2

Species scientific name: *Mullus barbatus*

Species common name: red mullet

Area: GSA 18 (Southern Adriatic Sea)

Years: 2014-2017

The results from the case study 2 are presented at the ICES WKBIOPTIM3 (2019) report in Section 5.3.

5.3 Case study 3

Species scientific name: *Mullus barbatus*

Species common name: red mullet

Area: GSA 22 (Aegean Sea)

Years: 2014, 2016

The results from the case study 3 are presented at the ICES WKBIOPTIM3 (2019) report in Section 5.5.

6. Conclusions

SampleOptim is a useful tool to define the number of individuals to sample by length class, and also to evaluate/test the effects that changes in the sampling stratification (based on Time Period, Port, Fleet and Sex) have on the age-length keys and on the maturity ogive.

7. Improvements required

The planning improvements to be included are: more quality indicators to define the optimum sample number by length class for the construction of age-length keys and the maturity ogive. The tool will also be improved with an output table where the optimum number, obtained from the different statistical methods applied, will be displayed.

8. GitHub links:

https://github.com/gonpatricia/SampleOptimRDBformat

https://github.com/ices-eg/wk_WKBIOPTIM3//SampleOptim