Workshop on Optimization of Biological Sampling (WKBIOPTIM 3)

Bitetto, Isabella; Brooks, Mollie Elizabeth; Craig, Jessica; Dubroca, Laurent; Fernandes, Ana C.; Gonçalves, Patricia; Håkansson, Kirsten Birch; Hilvarsson, Annelie; Lambert, Gwladys; Mantopoulou, Danai

Link to article, DOI: 10.17895/ices.pub.5647

Publication date: 2019

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
WORKSHOP ON OPTIMIZATION OF BIOLOGICAL SAMPLING (WKBIOPTIM 3)

VOLUME 1 | ISSUE 78
ICES Scientific Reports

Volume 1 | Issue 78

WORKSHOP ON OPTIMIZATION OF BIOLOGICAL SAMPLING (WKBIOPTIM 3)

Recommended format for purpose of citation:


Editors

Ana Cláudia Fernandes • Eirini Mantzouni

Authors

Isabella Bitetto • Mollie Elisabeth Brooks • Jessica Craig • Laurent Dubroca • Ana Cláudia Fernandes • Patrícia Gonçalves • Kirsten Håkansson • Annelie Hilvarsson • Gwladys Lambert • Danai Mantopoulou • Eirini Mantzouni • Katja Norén • Ioannis Ntokos • Nuno Prista • Dália Reis • Kostas Touloumis • Sofie Vandemaele • Julia Wischnewski
Contents

1 Introduction ..................................................................................................................... 1

1.1 Terms of Reference ...................................................................................................... 1

1.2 WKBIOPTIM 3 participants and agenda ................................................................. 1

1.3 Background to WKBIOPTIM 3 and report outline .................................................... 1

1.3.1 Overview of present WKBIOPTIM Optimization ToolBox .................................. 2

2 ToolBox description ....................................................................................................... 6

2.1 Sample Level Optimization (SampleLevelOptim) ...................................................... 6

2.1.1 Rationale of the analyses ...................................................................................... 6

2.1.2 Development/Improvements of the scripts during WKBIOPTIM3 ..................... 6

2.1.3 Next steps ............................................................................................................. 7

2.2 BioSimTool ................................................................................................................. 8

2.2.1 Rationale of the analyses ...................................................................................... 8

2.2.2 Quality indicators.................................................................................................. 8

2.2.3 Development/Improvements of the scripts during WKBIOPTIM3 ..................... 8

2.2.4 Next steps ............................................................................................................. 10

2.3 SampleOptim ............................................................................................................. 10

2.3.1 Rationale of the analyses ...................................................................................... 10

2.3.2 Quality Indicators.................................................................................................. 11

2.3.3 Development/Improvements of the scripts during WKBIOPTIM3 ..................... 12

2.3.4 Next steps ............................................................................................................. 13

2.4 SampleReferenceLevel ............................................................................................. 13

2.5 Multi-level analysis of sampling effort needed to produce estimates of catch numbers for stock assessment: SimPop ................................................................. 15

2.5.1 Rationale of the analyses ...................................................................................... 15

2.5.2 Development/Improvements of the scripts ........................................................ 16

2.6 SDTool ......................................................................................................................... 16

2.6.1 Rationale of the analyses ...................................................................................... 16

2.6.2 Quality indicators.................................................................................................. 17

2.6.3 Development/Improvements of the scripts during WKBIOPTIM3 ..................... 17

2.6.4 Next steps ............................................................................................................. 17

2.7 FishPi4WKBIOPTIM ................................................................................................. 17

2.7.1 Rationale of the analyses ...................................................................................... 17

2.7.2 Quality indicators.................................................................................................. 18

2.7.3 Development/Improvements of the scripts during WKBIOPTIM3 ..................... 18

2.7.4 Next steps ............................................................................................................. 18

2.8 Standardized notation for scripts ............................................................................. 19

2.9 Use of replacement (‘wr’) and without replacement (‘wor’) during simulations ...... 19

3 ToR b) Quality Indicators ............................................................................................ 20

3.1 Quality indicators used in WKBIOPTIM R toolbox .................................................. 20

4 ToR c) – ToolBox guidelines ....................................................................................... 21

4.1 Decision key for end users ....................................................................................... 21

5 Case studies .................................................................................................................. 25

5.1 SampleReferenceLevel Tool: Application in three case – studies............................ 25

5.1.1 Plaice in subarea 27.4.b, Quarter 4 (2018): German fleet length sampling data .... 25

5.1.2 Haddock in Division 27.7.e, Quarter 1 (2015): English fleet length sampling data.... 28

5.1.3 Plaice in Division 27.7.d, all quarters (2018): Belgian fleet length sampling data.... 31

5.2 Case study: Greece Purse seine (WKBIOPTIM2 Multi-level analysis & SDTool) ...... 35

5.2.1 WKBIOPTIM2 Multi-level analysis .................................................................... 35
Executive summary

The third Workshop on Optimization of Biological Sampling (WKBIOPIM3) convened to discuss practical aspects of optimization of sampling and provide a compilation of methods so that this information is well documented and organized for end users. New optimization algorithms were presented alongside developments and improvements of the work carried out in the earlier 2017 and 2018 workshops. Six different approaches/tools were presented: code developed under WKBIOPIM 1 and 2 for optimization at sample level (SampleLevelOptim), SampleOptim R-tool to optimize fish sampling for biological parameters, code for quantifying robustness of a length frequency distribution shape (SampleReferenceLevel), SDTool and BioSimTool as contributions from the STREAM Project to the optimization at sample-level and sampling design level, respectively, WKBIOPIM code used for optimization at sampling design level (SimPop). Additionally, data exploration and biological simulation tools developed under the framework of fishPi2 (FishPi4WKBIOpim). R-scripts for the different tools are available on the WKBIOPIM3 GitHub (https://github.com/ices-eg/wk_WKBIOPIM3).

Usability testing, including thorough instructions, were a focus of the workshop. As such, the group used a two-tiered system to evaluate the methods. The first subgroup tested the different scripts with their own data to see if they were of easy use. This first group also evaluated the documentation needed and whether it was clear how to interpret the results properly, how to decide on the best procedure according to their objectives, and how to compare results obtained using similar approaches. Based on this assessment, the second subgroup then cleaned and documented the code, discussing the approaches used in the different scripts and potential options for integration. In parallel, some participants developed a first draft of standardized notation aimed at harmonizing the documentation of simulation procedures used in the different codes, tested a new quality indicator for length frequency and developed scripts that demonstrate the effects of the common options of resampling ‘with replacement’ (wr) and ‘without replacement’ (wor) in the precision and bias of estimates. The workshop concluded with participants highlighting that specific documentation regarding quality indicators concepts and functions could also be very practical and useful for end users. Work of WKBIOPIM will continue intersessionally towards a final discussion of methodologies and results in 2020.
ii Expert group information

<table>
<thead>
<tr>
<th>Expert group name</th>
<th>Workshop on Optimization of Biological Sampling (WKBIOPTIM 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert group cycle</td>
<td>NA</td>
</tr>
<tr>
<td>Year cycle started</td>
<td>2019</td>
</tr>
<tr>
<td>Reporting year in cycle</td>
<td>1/1</td>
</tr>
<tr>
<td>Chairs</td>
<td>Ana Cláudia Fernandes, Portugal</td>
</tr>
<tr>
<td></td>
<td>Eirini Mantzouni, Greece</td>
</tr>
<tr>
<td>Meeting venue and dates</td>
<td>27-31 May 2019, Lysekil, Sweden (18 participants)</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Terms of Reference

The Third Workshop on Optimization of Biological Sampling (WKBIOPTIM 3) chaired by Ana Cláudia Fernandes (Portugal) and Eirini Mantzouni (Greece) met in Lysekil, Sweden, 27-31 May 2019 to:

a) **R-Toolbox**: Finalization and integrating the different developed scripts, including documentation

b) **Quality Indicators**: Discuss and conclude on a combination of indicators to evaluate the quality of data under different sample sizes, according to end users’ needs

c) Produce a guide for adequate use of sampling optimization procedures at national level, taking into account the results obtained in the analysis of the presented case studies (WKBIOPTIM 1 and WKBIOPTIM 2) and on the ongoing national experiences.

1.2 WKBIOPTIM 3 participants and agenda

The list of participants and the agenda for the workshop can be found in Annex 1 and Annex 2, respectively.

1.3 Background to WKBIOPTIM 3 and report outline

The WKBIOPTIM series aims to develop R-tools that evaluate if sampling effort for biological parameters (and associated resources) can be optimized without compromising the quality of final estimates. The need for those tools had been highlighted by several ICES EG’s (e.g. PGCCDBS 2012, PGDATA 2015, WKCOSTBEN 2016) that suggested that oversampling in lower stages of national sampling programs (e.g., number of trips, hauls within trips, fish within hauls) and inefficient sampling effort distribution may not be providing significant additional information on the population itself. In response to this, some national labs started developing statistical tools with the aim of analysing and optimize biological sample sizes by reducing on some clear-cut cases of excessive sampling at sample-level or increase when information collected is not sufficient. This work was presented at the first WKBIOPTIM and jointly developed by its participants thereafter.

All WKBIOPTIM R-scripts are based on the widely available RDB format for commercial sampling data. Scripts for conversion of DATRAS format to the RDB format have also been developed that allow some of WKBIOPTIM tools to run on research survey data. Overall, the main part of the preparation/development of the scripts has been made prior to the WKs themselves, with discussions and improvements being made during the WKs, and the final consolidation of the work (code and case studies) being carried out in the days after the WK. The R-scripts prepared/developed/improved in WKBIOPTIM’s keep being tested in different case-studies and the code is already being used by some institutes to analyse the improvements in sampling (e.g. by reducing oversampled species or evaluate how to increase sampling in other that are under sampled). The work performed during WKBIOPTIM’s has been presented in several groups and

ICES working groups (STECF, WGBIOP, WGCATCH, etc.) and have received positive feedback and good incentives to continue.

The third WKBIOPTIM aimed to compile the work prepared and presented in previous workshops and made it more readily available to users. New optimization algorithms were presented alongside developments and improvements of the work carried out in 2017 and 2018. The R scripts presented were made available to the participants through the GitHub. During the workshop, the practical work was divided between two subgroups: one for testing the scripts already available and the other for discussing, cleaning and documenting the scripts presented. In the first group participants tested the scripts with their own data to see if they are of easy use, to evaluate the type of documentation it will be needed, the interpretation of the results and how to evaluate them properly, how to decide on the best procedure according to their objectives, and to compare results obtained using similar approaches. The code developers group were discussing about the different approaches used by the several tools and about the possibility of integrating/adapting some of them. It was also proposed the preparation of a standardized notation to be used in all developed codes in order to harmonize and facilitate their use. At the end, participants highlighted that a specific documentation regarding quality indicators concepts and functions could also be very practical and useful for end users.

1.3.1 Overview of present WKBIOPTIM Optimization ToolBox

In summary, the newly or further developed R tools include:

**SampleLevelOptim:** This tool is designed for analysing the number of individuals that should be taken in each biological sample for the determination of length, age, sex and maturity. The algorithms have been developed during the previous meetings of WKBIOPTIM, and incorporate two different types of analyses: a) the original within-sample analysis developed under WKBIOPTIM1 (section 2.1) and b) the modelling approach developed under WKBIOPTIM2 (section 2.1). During the preparatory work for WKBIOPTIM3 the R-code previously developed was streamlined, optimized (through the use of parallel computing), and uploaded to GitHub. In addition, new vignettes were presented, discussed and tested during WKBIOPTIM3 by a set of colleagues less familiar with the optimization routine on their species of interest.

*Toolbox Description: Section 2.1*

*GitHub link:* [https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/sampleleveloptim](https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/sampleleveloptim)

*Case studies:* Section 4 of WKBIOPTIM1 and WKBIOPTIM2 reports

*Documentation:* not yet available (only case-study specific vignettes)

*Developer contact:* nuno.prista@slu.se

**BioSimTool:** The R tool BioSim, developed within STREAM WP3, aimed at providing support to the Commission and MSs for the realization of Multiannual Regional Work Programme and proposing a framework to "optimise" the sampling intensity for biological samplings. This tool is inspired to the scripts developed within WKBIOPTIM2 on the sampling optimization of biological variables (length, sex, maturity, age) at sample level. Although still working by trip, BioSim Tool derives a subset of indicators (CV and Earth Mover Distance) synthesizing the information from all the trips to derive an optimal number of biological measure (length, sex, maturity, age) to be collected on average by trip, on the basis of the relevant percentiles of the CV.
Some characteristics of the R tool and several examples of the outcome was presented and described.

**Toolbox Description:** Section 2.2

**GitHub link:** [https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/BioSim%20Tool%201.0](https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/BioSim%20Tool%201.0)

**Case studies:** Sections 5.3 and 5.6

**Documentation:** Annex 3C

**Developer contact:** bitetto@coispa.it

**SampleOptim:** The SampleOptim R-toolbox (SampleOptimRDB) aims to support user’s decision to determine the optimal sample size for age-length keys and maturity ogives. The SampleOptim R-tool was developed under the Portuguese sampling programme, the R-code was further adapted to the input data in the RDB format and presented during the WKBIOPTIM2. The setup for simulations allows the user to perform different types of scenarios taking into account a length stratification together with other possible stratifications (e.g. sex ratio, port and fleet). The simulation process works based on the input options at: sample level or quarter/semester/annual sampling level. 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.

**Toolbox Description:** Section 2.3

**GitHub link:** [https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SampleOptim](https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SampleOptim)
[https://github.com/gonpatricia/SampleOptimRDBformat](https://github.com/gonpatricia/SampleOptimRDBformat)

**Case studies:** Sections 5.3 and 5.6

**Documentation:** Annex 3A

**Developer contact:** patricia@ipma.pt

**SampleReferenceLevel:** The proposed general methodology to length sample optimization is distribution-free and was developed as an alternative to both parametric and bootstrap frameworks. The approach establishes quantitative objectives of length sampling not directly in the form of sample size determination, but in the form of selection of “information-sufficient”, adequate length frequency distribution (LFD). This is based on the concept of the reference or benchmark subsample. Reference subsample presents a theoretical “borderline scenario” or minimal adequate subsample, when a reasonably precise LFD comparing to original sample is still preserved. An iterative subsampling procedure based on certain criteria, returns a reference subsample as well quantified difference between original length sample and reference subsample, serving as a threshold.
Toolbox Description: Section 2.4
GitHub link: https://github.com/ices-eg/wk_WKBIOPIM3/tree/master/ADV
Case studies: Section 5.1
Developer contact: julia.wischnewski@thuenen.de

SimPop: Multi-level analysis of sampling effort needed to produce estimates of catch numbers for stock assessment

Toolbox Description: Section 2.2 in the WKBIOPTIM2 Report.
GitHub link: https://github.com/ices-eg/wk_WKBIOPIM3/tree/master/simpop
Case studies
Developer contact: laurent.dubroca@ifremer.fr

SDTool: This tool allows for the following analyses: (a) investigation on historical sampling data, (b) analyses for obtaining an “optimal” sampling size (in term of number of trips) based on the analyses of the CV of a given species, (c) simulations based on a fixed number of trips (that can be selected among the results from the optimization scripts) and, eventually, decreasing the number of length measurements, calculating the CV of the length distribution for each case study species, and (d) assessing the increase/decrease of CV by comparison to the past situation and implementing the calculation of EMD (Earth Mover’s Distance).

Toolbox Description: Section 2.6
GitHub link: https://github.com/ices-eg/wk_WKBIOPIM3/tree/master/SD%20Tool%202.04
https://github.com/aleligas/STREAM_MARE-2016-22/tree/Task-3.3-Sampling-Strategy-Optimization
Case studies: Sections 5.2 and 5.5
Documentation: Annex 3B
Developer contact: bitetto@coispa.it

FishPi4WKBIOPIM: The R package of fishpi2 WP3 which aimed at simulating length sampling was further developed during this Workshop. It moved away from sampling mean lengths to actually keeping track of the full length frequencies at the trip level. Further plots were developed to present these samples and some associated statistics. More raw outputs can now be extracted with the FishPi4WKBIOPIM package so that comparisons can be made with other available tools. The main wrapper function also now extracts a weighted CV, as presented by the BioSim Tool. As part of this workshop, all simulations outputs were documented in order to increase the transparency and usefulness of the framework and further examples were worked through in details to help with increasing the flexibility of the framework as well as debugging.
Toolbox Description: Section 2.7

GitHub link: https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/FishPl2WP3_BioSims

Case studies: Sections 5.4 and 5.5

Documentation: Annex 3D

Developer contact: gwladys.lambert@cefas.co.uk; kih@aqua.dtu.dk

**LanAge_opt:** 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. This tool is largely based on the tool during WKBIOPTIM3: Simulation on the number of age readings per length class (LC) and effect on the age structure of landings. It should be noted that this approach was finalised 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.

GitHub link: https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/lanAge_opt

Case studies: Section 5.6

Documentation: Annex 7

Developer contact: emantzo@inale.gr
2 ToolBox description

2.1 Sample Level Optimization (SampleLevelOptim)

2.1.1 Rationale of the analyses

See WKBIOPTIM 1 report, Section 2.1.1.

2.1.2 Development/Improvements of the scripts during WKBIOPTIM3

During the preparatory work for WKBIOPTIM3 the R-code previously developed was streamlined, optimized (through the use of parallel computing), and uploaded to GitHub. These vignettes now incorporate two different type of analyses: a) the original within-sample analysis developed under WKBIOPTIM1 (section 2.1) and b) the modelling approach developed under WKBIOPTIM2 (section 2.1). (see Figure 2.1.2.1)

![Documents library]

Figure 2.1.2.1 - Within-sample vignettes tested in WKBIOPTIM3

The new vignettes were presented, discussed and tested during WKBIOPTIM3 by a set of colleagues less familiar with the optimization routine. Outputs of one of these tests - run on samples of *Sprattus sprattus* – is displayed in Figure 2.1.2.2.
Figure 2.1.2.2 - Example graphical output of within-sample optimization run on a set of samples of Sprattus sprattus.

The work carried out during WKBIOPTIM3 indicates that the code and vignettes provide a suitable basis for more widespread use. For that to happen it will be useful if an R-package is built. To achieve that, a set of aspects that need development were identified, namely:

- Creation of a function for checking format of input data
- Generation of functions for main parts of the code (part of the information currently present in the vignettes should be incorporated into help-pages for these functions)
- Some compatibility issues between OS (Linux, Windows) and between software (RStudio, Notepad++) need to be addressed and code made general as much as possible.
- Better instructions need to be given on the use of auxiliary input files (e.g., variable_table.csv)
- More detailed explanations on how to interpret the main outputs are needed

Additional paths for development were also highlighted, namely interest in seeing previous options for simulation of different sampling strategies (e.g., stratification by length) re-introduced in the code. Additionally, even though it is not the focus of the current analyses, participants expressed the wish that graphical outputs displaying results obtained at quarter and annual level (i.e., after pooling re-samples with suggested sample sizes) are also incorporated in the code.

2.1.3 Next steps

The work towards incorporating the feed-back of WKBIOPTIM3 into the R-scripts and developing a final R-package will continue intersessionally, with developments expected to be presented during WBIOP4M. The work done on other simulation tools (Annex 3) provides a set of possible alternatives for the full documentation of functions and vignettes.
2.2 BioSimTool

GitHub link: https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/BioSim%20Tool%201.01

Case studies: Sections 5.3 and 5.6

Documentation: Annex 3C

Developer contact: bitetto@coispa.it

2.2.1 Rationale of the analyses

The BioSim Tool is based on resampling techniques of the commercial data and allows to perform the following analyses:

- on the length frequency distributions at sample level (by trip or by trip and commercial category) to find the minimum number of fish to be measured to obtain representative samples of the sampled population avoiding oversampling;

- on the other biological variables (age, sex and maturity) at sample level (by trip or by trip and commercial category), varying the number of individuals for which the biological measures are taken and estimating the sex-ratio at length, the maturity-at-length and the age-length keys (ALK) for each resampling.

2.2.2 Quality indicators

MWCV%

It is the precision of the number of observations in each length class of a random sample that can be estimated by assuming a multinomial distribution. If the precision in each length class is in the form of a coefficient of variation (CV), an overall measure of precision can be obtained by weighting each CV by the number of fish in each length class. This mean weighted CV (MWCV) provides a description of the precision over the entire size range in a length frequency distribution (Gerritsen and McGrath, 2007). Under the assumption of a multinomial distribution, the standard deviation ($\sigma_i$) of the number of fish in a specific length class $i$ can be estimated by:

$$\sigma_i = \sqrt{np_i(1-p_i)}$$

(1)

where $n$ = the total number of fish in the sample and $p_i$ = the proportion of the sample with length $i$.

The coefficient of variation (CV) of the number of fish at length $i$, is given by:

$$CV_i = \frac{\sigma_i}{np_i}$$

(2)

and the mean weighted coefficient of variation (MWCV) is given by:

$$MWCV = \sum p_i CV_i = \frac{\sum \sigma_i}{N}$$

(3)
Earth Mover’s Distance

Earth Mover’s Distance can be formulated and solved as a transportation problem. Suppose that several suppliers, each with a given amount of goods, are required to supply several consumers, each with a given limited capacity. For each supplier-consumer pair, the cost of transporting a single unit of goods is given. The transportation problem is then to find a least-expensive flow of goods from the suppliers to the consumers that satisfies the consumers’ demand. Similarly, here the problem is to transform one length distribution (LFDscenario_n) to another (LFDbaseline) with minimum work done, then the lower is the EMD, the greater the similarity between the two distributions.

A function from the R package `emdist` developed by Urbanek (2015) and available on the R-CRAN repository is embedded in the BioSim Tool for the estimation of the EMD.

The statistics and indicators to identify the “optimal” number of individuals to be measured “at sample level” (i.e. per trip) calculated per each resampled trip are the following (ICES WKBIOPTIM, 2017):

- mean, median, minimum, maximum and CV of the lengths and number of sampled classes;
- MWCV (mean weighted coefficient of variation) of the sample;
- number of sampled length classes and number of modes (smooth).

CV of sex ratio at length

For the optimization of sex data, for each simulation the following outputs are produced:

- sex ratio at length calculated as the ratio between the number of females to the sum of the number of females and males in each length class;
- total CV of the sex ratio in order to assess the trend of the precision when varying the number of sexed individuals.

The formula used to estimate the total CV is:

\[
CV = \sqrt{\frac{\sum_{l} (SR(l) - \bar{SR})^2 \times (1 - SR(l))}{\sum_{l} (N(l) - 1) \times (1 - SR(l))}} \times 100
\]

where \(N(l) = M(l) + F(l)\) the number of individuals at length \(l\), \(M(l)\) are the males of class \(l\), \(F(l)\) are the females of class \(l\), \(SR(l) = \frac{F(l)}{M(l) + F(l)}\) the sex ratio at length \(l\), and \(SR = \frac{\sum_{l} F(l)}{\sum_{l} M(l) + \sum_{l} F(l)}\) the overall sex ratio.
CV of proportion of matures at length

For the optimization of the number of maturity measures, the formula used to estimate the total CV is:

\[
CV = \sqrt{\frac{\sum p(l)(1-p(l))}{N-1}} \times 100
\]  

(5)

Where \( p(l) = \frac{mat(l)}{imm(l)+mat(l)} \) the proportion of matures at length \( l \), \( imm(l) \) are the immatures at length class \( l \), \( mat(l) \) are the matures at length class \( l \), \( N = \sum mat(l) + \sum imm(l) \) the total number of individuals considered and \( p = \frac{\sum mat(l)}{\sum mat(l)+\sum imm(l)} \) is the overall proportion of mature individuals.

CV of age-length key

For the optimization of the number of age readings by length class, the precision in terms of CV related to the relationship between length and age is calculated through the analytical study of the variance, following the method based on the multinomial distribution proposed by Baird (1983).

2.2.3 Development/Improvements of the scripts during WKBIOPTIM3

During the WKBIOPTIM3, the scripts were not modified.

2.2.4 Next steps

The development of an R tool including all the scripts implemented in the BioSim Tool would be also important to allow a wider availability to the method and to ease the application to other case studies.

2.3 SampleOptim

GitHub link: https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SampleOptim

Case studies: Sections 5.3 and 5.6

Documentation: Annex 3A

Developer contact: patricia@ipma.pt

2.3.1 Rationale of the analyses

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 (ALKs) and maturity ogives (MO). The simulation process works based on the input options at:
sample level or quarter/semester/annual level. The dataset used should represents the “whole” available data, and the simulations are based on subsampling without replacement by default.

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 different types of scenarios taking into account a length stratification together with other possible stratifications:

- temporal stratification (annual, semester and quarter);
- sex stratification, by defining the sex ratio 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.

The main aim of SampleOptim is to evaluate a reduction in the number of samples collected to estimate the biological parameters used on the assessment. Although, in cases were the actual sample size is not enough, the functions in the algorithms will not converge which could indicate some bias on the original data sampling.

### 2.3.2 Quality Indicators

In order to allow the comparison of the different datasets, resulting from each simulation scenario, a series of quality indicators are estimated to evaluate the different options and help on the decision process of finding the optimum sample size by length class. Since these data will be used as input into assessment, for the ALKs and the MO parameters, the comparison of the following estimates between the original data and the simulations data subsets are used: mean length-at-age, mean age-at-length, the parameters of the von Bertalanffy growth model (L_{inf}, t_0 and k) (section 2.3.2.1) and the maturity ogive parameters (L_{25}, L_{50} and L_{75}) (section 2.3.2.2).

Besides the former quality indicators, the root mean squared prediction error (RMSPE) (section 2.3.2.3), mean squared prediction error (MSPE) (section 2.3.2.4) and the mean average percentage error (MAPE) (section 2.3.2.5) are also used.

#### 2.3.2.1 von Bertalanffy growth model (VBGM)

It is the most widely used growth model and is especially important in fisheries studies, von Bertalanffy derived this equation in 1938 from simple physiological arguments. The model assumes that the rate of growth of an organism declines with size so that the rate of change in length, l, may be described by:

\[
\frac{dl}{dt} = k(L_{inf} - l)
\]

where: \(t\) is time, \(l\) is length (or some other measure of size), \(k\) is growth rate, \(L_{inf}\) is the asymptotic length at which growth is zero.

By integrating the previous equation, it becomes:

\[
l t = L_{inf} (1 - e^{-k(t-t_0)})
\]

The parameter \(t_0\) is included to adjust the equation for the initial size of the organism and is defined as the time at which the organisms would have had zero size. Thus to fit this equation, is necessary to fit the 3 parameters (\(L_{inf}, k\) and \(t_0\)) by nonlinear regression.
2.3.2.2 Maturity ogive (MO)
A generalized linear model (GLM) with a binomial error distribution and a logit link was fitted to the proportion of fish mature by length class. The model has the general form:
\[
\log(E[m]) = a + bl
\]
where \(a\) and \(b\) are the intercept and slope of the ogive.

Estimates of the length at which 50% of the individuals were mature (L50) and of the slope of the ogive at L50 were derived from the model parameters as:
\[
L50 = \frac{-b}{a}
\]
A similar procedure as applied to estimate L25 and L75.

2.3.2.3 Root mean squared prediction error (RMSPE)
The RMSPE of a parameter estimator is the square-root of the mean squared error (MSPE) of the estimator. In symbols, if \(\hat{\theta}\) is an estimator of the parameter \(\theta\), then
\[
\text{RMSPE}(\theta) = \sqrt{E[\hat{\theta} - \theta]^2}
\]
The RMSPE of an estimator is a measure of the expected error of the estimator. The units of RMSPE are the same as the units of the estimator.

2.3.2.4 Mean squared prediction error (MSPE)
The MSPE is the mean square error of the estimator. In symbols, if \(\hat{\theta}\) is an estimator of the parameter \(\theta\), then
\[
\text{MSPE}(\theta) = E[\hat{\theta} - \theta]^2
\]

2.3.2.5 Mean absolute percentage error (MAPE)
The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), measures the accuracy of a method for constructing fitted time series values in statistics. It calculates the mean absolute percentage error (Deviation) function for the forecast and the eventual outcomes. The mean absolute percentage error (MAPE) is defined as:
\[
\text{MAPE} = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{x_i - \hat{x}_i}{x_i} \right|
\]
where \(x_i\) is the actual observations time series, \(\hat{x}_i\) is the estimated or forecasted time series and \(N\) is the number of non-missing data points.

2.3.3 Development/Improvements of the scripts during WKBIOP-TIM3
The “00.Data conversion and preparation.R” R-script has included during the WKBIOPTIM3 to convert the CA.csv RDB data file into the required dataset format (see SampleOptim Documentation – Annex 3-A3).

The initial/input parameters table file (input_params.csv) has been created and added to “2_Simulations_RDB.R”, as the setup parameters used when running the simulation data operational model (“sample_selection_function_RDB.R”).
2.3.4 Next steps

The improvements planned 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.

2.4 SampleReferenceLevel

GitHub link: https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/ADV

Case studies: Section 5.1

Developer contact: julia.wischnewski@thuenen.de

The proposed general methodology to length sample optimization was described in Wischnewski et al (in prep) and presented at WKBIOPTIM3. It is distribution-free and it was developed as an alternative to both parametric and bootstrap frameworks.

The approach establishes quantitative objectives of length sampling not directly in the form of sample size determination, but in the form of selection of “information-sufficient”, adequate length frequency distribution (LFD). This is based on the concept of the reference or benchmark subsample. Reference subsample presents a theoretical “borderline scenario” or minimal adequate subsample, when a reasonably precise LFD comparing to original sample is still preserved. An iterative subsampling procedure based on certain criteria, returns a reference subsample as well quantified difference between original length sample and reference subsample, serving as a threshold. We have called this threshold admissible dissimilarity value (ADV). The criteria defined in procedure include both statistical arguments (which characteristics of a distributional shape are important) and biologically practical aspects (how many measurements/hauls/trips we are able to afford, depending on species, area, fleet etc.). For analysis, the data in the Regional Data Base (RDB) Exchange Format raised to haul level were utilized.

Statistical aspects are based on concept of robust mode/antimode of distributional shape. Formally, length class is determined as a mode (antimode) if adjacent length classes exhibit lower (upper) frequency values (i.e. local maxima/minima). The following definition concerns the meaning of significantly distinct, robust modes and antimodes in our context.

**Definition 1.** Let $\bar{M} = (M_1, M_2, \ldots)^T$ be modes and $\bar{A} = (A_1, A_2, \ldots)^T$ be antimodes of some LFD with bandwidth 1 cm, and $\bar{M} \text{smoothed} = (M_1^\Delta, M_2^\Delta, \ldots)^T$ be modes and $\bar{A} \text{smoothed} = (A_1^\Delta, A_2^\Delta, \ldots)^T$ be antimodes of the same LFD with selected bandwidth $\Delta > 1$ cm, where $\Delta = \Delta (\text{max species length})$.

We define a mode $M_i \in \bar{M}$ as a robust mode, if

1. $M_i^\Delta \leq M_i < M_i^\Delta + \Delta$
2. $M_i = \max_{(M_1, M_2, \ldots) \in [M_i^\Delta, M_i^\Delta + \Delta]} (M_1, M_2, \ldots)$
3. $f(M_i) > 0.01 \cdot \max(f(M_1), f(M_2), \ldots)$.

In the same way, an antimode $A_j \in \bar{A}$ is a robust antimode, if

1. $A_j^\Delta \leq A_j < A_j^\Delta + \Delta$
2. $A_j = \min_{(A_1, A_2, \ldots) \in [A_j^\Delta, A_j^\Delta + \Delta]} (A_1, A_2, \ldots)$. 
In other words, despite length classes smoothing with bandwidth values $\Delta > 1$ cm, the robust modes and antimodes are really there and not suspected sampling artifacts. For the case of existence of two or more original modes/antimodes within $\Delta$-smoothed length class, a dominating one (i.e. maximal/minimal) will be selected as a robust mode/antimode. The choice of $\Delta$ depends on maximum length of certain species. Moreover, modes having too low frequencies (less than 1% of maximal frequency) cannot be considered as robust ones. Of course, depending on requirements, all modes/antimodes can be considered as robust ones.

The next definition provides formal requirements of statistical-biological similarity between LFDs of original and subsampled data sets.

**Definition 2.** Let $\vec{M} = (M_1, M_2, \ldots)^T$ and $\vec{m} = (m_1, m_2, \ldots)^T$ be robust modes and $\vec{A} = (A_1, A_2, \ldots)^T$ and $\vec{a} = (a_1, a_2, \ldots)^T$ be robust antimodes of the original and reduced samples LFDs, respectively. We define those as similar, if:

1. they have the same number of robust modes and antimodes, i.e. $\dim(\vec{m}) = \dim(\vec{M})$ and $\dim(\vec{a}) = \dim(\vec{A})$;
2. for each corresponding pair $m_i, M_i$ and $a_j, A_j$:
   $|m_i - M_i| \leq \varepsilon$ and $|a_j - A_j| \leq \varepsilon$, where $\varepsilon = \varepsilon (\text{max species length})$
3. amplitudes ratio $\frac{|m_i - a_j|}{|M_i - A_j|} \geq \theta$, where $i \in \{j; j + 1\}$, $0 < \theta \leq 1$.

Informally, this definition says that subsampled data set has to preserve a structure, or specific patterns, of the original data set, revealing the same number of robust critical values (modes and antimodes) as well as keeping almost constant differences in amplitudes between adjacent extreme values. Locations of critical values for larger species have not to be exact and can vary in some small interval defined by parameter $\varepsilon$. Our assumption is that if conditions (1)-(3) are satisfied, two samples are indistinguishable in statistical-biological sense.

Note that conditions 1-3 might be extended for proposed subsampling procedure. We introduce the following additional (optional) conditions.

4. One can also set, if desired, a value $\gamma$ indicating a minimal required number per length class in subsample. This can be a fixed number of individuals in each length class (e.g. scalar value $\gamma = 1000$ fishes), or relative number (e.g. a percentage of the original number in each length class, so vector value $(\gamma_j)^L_{j=1}$, where $L$ is a number of 1 cm length classes). We can say that parameter $\gamma$ reflects the requirements of official national sampling program in a certain sense.

5. It is not necessary to be focused on all length classes $l_j$, $j = 1, 2, \ldots, L$. We can choose important ones for us, or target, length classes $l^t = \{l_j\}$, $j \in I \subseteq \{1, 2, \ldots, L\}$, providing the most significant information about LFD shape and ignore other ones (e.g. very small or very large fishes). Then the subsampling process as well as ADV will be defined only on set $l^t$.

The following algorithm scheme is proposed:

1. Use as basic input data from the standard RDB data with length rounded to 1 cm.
2. Select bandwidth $\Delta$ and important length classes $l^t$ if desired (condition 5), identify corresponding robust modes and antimodes (critical values) in the original sample under $\Delta$ on set $l^t$.
3. Set remaining parameters $\{\theta, \gamma, \varepsilon\}$.
4. Remove one length measurement from each subsampling and see whether conditions 1-4 are satisfied. If yes, repeat the step. If no, go back to the previous subsample and stop.

Thus, a subsample obtained as result of this procedure, presents a reference subsample, which ensures an evidence of sufficient information about distributional shape and represents important length groups well, comparing to the original sample.

The ADV is defined as a value showing difference between original sample and constructed reference subsample. In our approach we employ a well-known Manhattan or 1-Wasserstein distance, or $L_1$-distance

$$L_1(F, G) = \sum_{j=1}^{I} |F_{lj} - G_{lj}|,$$

where $F$ and $G$ are empirical cumulative distribution functions (CDFs) of original and any subsampled data sets, respectively, $F_{lj}$ and $G_{lj}$ are their values in length class $l_j \in l^I$. Thus ADV will be determined as $L_1(F, G_{ref})$, where $G_{ref}$ is a reference subsample described in previous subsection. It defines a threshold to complete our framework to decide on acceptable and unacceptable dissimilarities between LFDs when reducing sampling effort.

A summary of future work planned follows:

- consideration of regional sampling programs as a practical target
- testing other statistical distances for ADV computation, possibly creation of a new hybrid distance.
- improving of the iteration scheme (e.g. computational efficiency, general formalization, extension towards more sophisticated effort optimization variants, etc.).
- inclusion of parameters or requirements defining the sampling costs or/and environment factors.
- trying to differentiate between target and incidental catch, as well as between at-sea and onshore sampling. Then definition of a reference subsample by a set of considered parameters might make it more specific.

2.5 Multi-level analysis of sampling effort needed to produce estimates of catch numbers for stock assessment: SimPop

GitHub link: https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/simpop

Case studies: see Section 5.2 and the WKBIOPTIM2 Report (Section 5.1)

2.5.1 Rationale of the analyses

The rationale of the analyses follows the framework presented in the WKBIOPTIM 1 and 2 reports.

Data collected in the EU-MAP framework are aimed first to provide stock information for the assessment working groups. National monitoring program and sampling plan have to be tailored to fit this need. Consequently, optimization of the sampling plan has to follow the same constraint: to provide the estimates requested by the end-users.

In this framework, analyses were carried out to test how sampling plan changes can affect population estimates. The estimator used in this exercise was the length distribution estimate. Compared to age estimates or discards estimates, this estimate is the most simple to provide to assessment working groups. Moreover, measuring length is the archetypal activity in field work
related to commercial fisheries and data are widely available in the national databases. Starting from the existing sampling database, simulations were used to modify the sampling plan. The simulation process includes changes in the sampling effort by decreasing the amount of samples or measurements and then computing new length distribution estimates using the modified sampling database. The new estimates were compared with the original length distribution to assess the effect of reducing samples on the estimates.

During the first WKBIOPTIM workshop, impact on length distribution were assessed at a stock level. During this second WKBIOPTIM workshop, impact on length distribution was assessed for a given area for all the main species included in the sampling plan. The main objective was to assess sampling plan optimization in a concurrent sampling framework. During the WKBIOPTIM 3 workshop, some improvements were made to evaluate the impact of sampling reduction on the age distribution estimates.

### 2.5.2 Development/Improvements of the scripts

The scripts developed during the past two WKBIOPTIM were used during this one. No major modification was introduced in the scripts used before, and no improvement were made on the case studies.

### 2.6 SDTool

**GitHub link:** [https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SD%20Tool%202.04](https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SD%20Tool%202.04)

**Case studies:** Sections 5.2 and 5.5

**Documentation:** Annex 3B

**Developer contact:** bitetto@coispa.it

### 2.6.1 Rationale of the analyses

In the project MARE/2014/19 Med&BS a Sampling Design tool based on bootstrap technique (Deliverable 2.5 by F. Gontrand and T. Rouyer) was developed in R language to explore the consequences in terms of variation of precision (Coefficient of Variation (CV)) of several changes in the current sampling design (e.g. trips). This was achieved through the resampling of historical data. The underlying statistical principle is that the value of CV decreases with the increase of the number of sampling units, defining a curve. In the Sampling Design tool, the part of the curve where the tangent changes and begins to flatten (i.e. the curvature range) is considered as a suitable trade-off between precision and sampling effort and proposed as “optimal” sample size (in terms of number of trips).

In the STREAM project (Strengthening Regional cooperation in the area of fisheries biological data collection in the Mediterranean and Black Sea) the SD Tool R scripts developed in the previous grant were generalized to allow the optimization also by GSA, by area (e.g. combination of GSAs), country and gear (level 4). Moreover, new scripts to compare and evaluate the precision of different sampling stratifications were developed aimed at supporting the identification of the best trade-off between sampling effort (in terms of number of trips) and precision (CV) by including in the optimization process several inputs from expert knowledge (i.e. a constraint of at least one trip by month).
2.6.2 Quality indicators

The CV per length class and total are calculated on each bootstrapped sample according to Vigneau and Mahevas (2007), by means of the functions implemented in the COSTdbe library (version 1.4).

For the EMD definition, see the definition on the BioSim quality Indicators paragraph.

2.6.3 Development/Improvements of the scripts during WKBIOPTIM3

During the WKBIOPTIM3 the scripts were not modified.

2.6.4 Next steps

The inclusion of more indicators would undoubtedly be beneficial to provide additional information on the impact of alternative sampling designs on the LFDs.

The development of an R tool including all the scripts implemented in the SDTool would be also important to allow a wider availability to the method and to ease the application to other case studies.

2.7 FishPi4WKBIOPTIM

GitHub link: https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/FishPi2WP3_BioSims

Case studies: Sections 5.4 and 5.5

Documentation: Annex 3D

Developer contact: gwladys.lambert@cefas.co.uk; kih@aqua.dtu.dk

2.7.1 Rationale of the analyses

Fishpi2 was an EU-funded project that brought together scientists from 10 Member States with the aim to strengthen regional cooperation on the collection of biological fisheries data in EU waters and provide guidance on implementation of regional sampling plans. The project ended mid-2019 and it comprised six work packages. The tool presented here was initially developed under WP3, which aimed to develop a framework to help assess and compare the performance of potential regional sampling plans for commercial fisheries. In that sense, the framework was built with the ability to simulate stratified sampling strategies either on-shore or at-sea, with a given sampling effort per stratum, and one or two-stage sampling designs (the fishing trip being the smallest sampling unit either way; two-stage would be, for example, site-day as primary sampling unit (PSU) and trip as secondary sampling unit (SSU)).

Full datasets of lengths or age distributions, which are the main biological variables of interest annually for stock assessment purposes, do not exist at the population level. Only what is currently sampled can be used in simulations based on real data, either raw as such or raised at the population level following some rules based on the sampling design and ad hoc information on the populations of interest (e.g. grouping of gear types).

One thing that exists is the full population of landings in weight and at the trip level (with exception of the smaller vessels for which information may be incomplete or of any other specific activity that may be missed in officially recorded data). The main framework in fishpi2 WP3 thus
used estimated landed weight as a proxy to measure the effect of a sampling design on actual parameters of interest (length/age distribution).

A more specific data call under WP3 focused on getting length sample data at the trip level for a restricted number of species and areas. As a smaller exercise compared to the full set of simulations on landings, the framework was adapted to estimate mean length rather than total landings. These simulations assumed the sample dataset was the “full population”, i.e. no raising was attempted as part of the project, although it was recognised that this was a limitation.

In the package presented here, the user must provide a full dataset of the population to sample from (population of interest), a covariate for stratification (if required), the sampling frame (the variable representing the primary sampling unit, PSU), the secondary sampling unit (SSU variable) (no tertiary unit, or TSU, option included yet), the effort by stratum (number of PSUs to sample), the effort at PSU level (number of SSUs to sample) and the domains of interest. Domains do not affect the sampling; they are just entities of interest to report metrics over (e.g. combination species-area as a proxy for stock). By, for example, varying the effort (e.g. number of trips sampled per strata) or changing the stratification scheme, the user can compare the efficiency of a range of sampling designs.

### 2.7.2 Quality indicators

Graphical methods were developed as part of the main FishPi2 WP3 framework to support sampling design selection. These focused mainly around the point estimate and its confidence interval compared to the true value and on the relative standard error (RSE), defined as standard deviation over mean of estimates (across simulation replicates). RSE is expected to decrease as sampling effort and sampling efficiency increases. During WKBIOPTIM3 the calculation of the MWCV was added.

### 2.7.3 Development/Improvements of the scripts during WKBIOPTIM3

During the workshop and for the present report, the R package of FishPi2 WP3 which aimed at simulating length sampling was further developed. It moved away from sampling mean lengths to actually keeping track of the full length frequencies at the trip level. Further plots were developed to present these samples and some associated statistics. More raw outputs can now be extracted with the FishPi4WKBIOPTIM package so that comparisons can be made with other available tools. The main wrapper function also now extracts a weighted CV, as presented by the BioSim Tool. As part of this workshop, all simulations outputs were documented in order to increase the transparency and usefulness of the framework and further examples were worked through in details to help with increasing the flexibility of the framework as well as debugging.

### 2.7.4 Next steps

As became evident while further developing this R package for WKBIOPTIM3, this framework has a lot of potential for further development in line with researching improvements that can be made to regional, or even more local, sampling plans. One of the assets of this framework is that the simulations are run from one main function that is relatively short, efficient and straightforward to read through in order to understand what the package does. One drawback is that the outputs could be shared with more clarity and ease of extraction for an external user.
At this stage, the package could benefit from further testing, using different case studies and producing more standard outputs to be more easily compared with other tools where appropriate. A list of development requirements or ideas is also provided in the package documentation section of this report (Annex 3D). These include that the package should be developed to be able to simulate concurrent sampling options (sampling one species at a time for now) and to have different/more flexible levels of sampling: the haul level for example, and the number of lengths data, or the potential to sample age-at-length data.

A summary of future work planned follows:

- check formatRDB2FishPi function to facilitate the formatting of standard datasets
- provide statistical outputs based on length frequency (only exploratory for now)
- make extracting outputs more intuitive
- set option for concurrent sampling (for now it conducts simulations one species at a time)
- make sure it can work with raised data
- run further checks to make sure it does as intended with stratification and with two-stage sampling (all examples worked up with single stage sampling until now)
- add hauls to the level of sampling – at the moment fishing trip is the minimum possible sampling unit and all fish from that trip, available in the dataset, is accounted for (could consider defining number of fish to sample too)
- add a level of sampling to sample age data within length bins (or sample ages directly rather than lengths)
- in general, make the sampling units and their hierarchy more flexible
- add possibility for including finite population correction in the calculation -most relevant when simulating on sample data only as this package currently does make the specification of domains and strata more flexible/generic

2.8 Standardized notation for scripts

A subgroup within WKBIOPTIM3 developed a proposal that aims to standardize the mathematical notation used in simulations, facilitating communication and documentation of the code. The proposal is displayed in Annex 6 and will be further refined intersessionally towards a final discussion during WKBIOPTIM4.

2.9 Use of replacement (‘wr’) and without replacement (‘wor’) during simulations

A subgroup within WKBIOPTIM3 developed code to examine the effects of re-sampling strategy with replacement (i.e., bootstrap) and without replacement (more physically tangible reality for lab personnel) on precision and accuracy of estimates. A set of preliminary results was presented during the meeting. Work will continue intersessionally with expected discussion during WKBIOPTIM4.
3 ToR b) Quality Indicators

3.1 Quality indicators used in WKBIOPTIM R toolbox

The quality indicators used in the various tools have been described in Section 2 and Annex 3 under each approach. An overview table is presented below:

<table>
<thead>
<tr>
<th>QI</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth Mover Distance (EMD)</td>
<td>BioSimTool</td>
</tr>
<tr>
<td></td>
<td>Simpop</td>
</tr>
<tr>
<td></td>
<td>SDTool</td>
</tr>
<tr>
<td>Mean-weighted CV of frequency (MWCV)</td>
<td>SampleLevelOptim</td>
</tr>
<tr>
<td></td>
<td>simpop</td>
</tr>
<tr>
<td></td>
<td>BioSimTool, FishPi4WKBIOPTIM</td>
</tr>
<tr>
<td>CV of the mean</td>
<td>SampleLevelOptim</td>
</tr>
<tr>
<td></td>
<td>SDTool</td>
</tr>
<tr>
<td></td>
<td>simpop</td>
</tr>
<tr>
<td>CV of ALK, CV of proportion of matures, CV of sex ratio</td>
<td>BioSimTool</td>
</tr>
<tr>
<td>Relative Standard Error (RSE)</td>
<td>FishPi4WKBIOPTIM</td>
</tr>
<tr>
<td>Number of size/age/sex classes sampled</td>
<td>SampleLevelOptim</td>
</tr>
<tr>
<td></td>
<td>BioSimTool</td>
</tr>
<tr>
<td>Number of modes identical with the modes of original sample before/after smoothing</td>
<td>SampleLevelOptim</td>
</tr>
<tr>
<td></td>
<td>BioSimTool</td>
</tr>
<tr>
<td>Two sample t-test for equality of means</td>
<td>SampleLevelOptim</td>
</tr>
<tr>
<td>Two sample Kolmogorov-Smirnov test</td>
<td>SampleLevelOptim</td>
</tr>
<tr>
<td>Robustness of modes/antimodes and admissible dissimilarity value</td>
<td>SampleReferenceLevel</td>
</tr>
<tr>
<td>Root mean squared prediction error (RMSPE)</td>
<td>SampleOptim</td>
</tr>
<tr>
<td>Mean squared prediction error (MSPE)</td>
<td>SampleOptim</td>
</tr>
<tr>
<td>Mean absolute percentage error (MAPE)</td>
<td>SampleOptim</td>
</tr>
<tr>
<td>von Bertalanffy growth model (VBGM) parameters</td>
<td>SampleOptim</td>
</tr>
<tr>
<td>Maturity ogive (MO)</td>
<td>SampleOptim</td>
</tr>
</tbody>
</table>
4 ToR c) – ToolBox guidelines

4.1 Decision key for end users

In the Table below the main characteristics and features of the tools developed and/or applied during WKBIOPTIM3 are presented, to assist the user to select the right one(s) for their optimization analyses.

<table>
<thead>
<tr>
<th>VARIABLE(S) OF INTEREST (what the precision is estimated for)</th>
<th>SDtool</th>
<th>BioSim Tool</th>
<th>Fishpi4WKBIOPTIM</th>
<th>SimPop</th>
<th>SampleOptim</th>
<th>SampleLevelOptim</th>
<th>SampleReferenceLevel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landings (or discards or catch)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Mean length</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Length distribution</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mean age at length</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No, but can be extended</td>
<td>No</td>
</tr>
<tr>
<td>Age distribution</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No, but can be extended</td>
</tr>
<tr>
<td>Sex ratio</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Maturity at age</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Inference on sample [1] or population [2] data</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SAMPLING LEVELS</td>
<td># Trip</td>
<td># Trip/haul</td>
<td>Fish</td>
<td># Trip</td>
<td># vessel/Trip</td>
<td># site-day/ Trip</td>
<td>Trip/haul</td>
</tr>
<tr>
<td>1: Single level unstratified</td>
<td>1, 2</td>
<td>1</td>
<td>1, 2</td>
<td>3</td>
<td>1, 2</td>
<td>1, 2</td>
<td>1</td>
</tr>
<tr>
<td>2: Single level stratified</td>
<td>1, 2</td>
<td>1</td>
<td>1, 2</td>
<td>3</td>
<td>1, 2</td>
<td>1, 2</td>
<td>1</td>
</tr>
<tr>
<td>3: Multi-level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRATIFICATION OPTIONS</td>
<td>SDtool</td>
<td>BioSim Tool</td>
<td>Fishpi4WKBIOPTIM</td>
<td>SimPop</td>
<td>SampleOptim</td>
<td>SampleLevelOptim</td>
<td>SampleReferenceLevel</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------</td>
<td>-------------</td>
<td>-------------------</td>
<td>--------</td>
<td>-------------</td>
<td>-------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Technical (eg metier)</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>yes (choice of specific metiers)</td>
<td>yes (choice of specific metiers)</td>
</tr>
<tr>
<td>Spatial (eg GSA)</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>yes (choice of specific ICES area)</td>
<td>yes (choice of specific ICES area)</td>
</tr>
<tr>
<td>Temporal (eg year)</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>yes (choice of specific year and quarter)</td>
<td>yes (choice of specific year and quarter)</td>
</tr>
<tr>
<td>other (specify)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Vessel parameters (length, GT...)</td>
<td>Sexratio, length</td>
<td>any biological variable (e.g., length, sex)</td>
<td>choice of specific sampling type</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CONCURRENT SAMPLING</th>
<th>1: all species in sample</th>
<th>2: can select which species to sample</th>
<th>3: can select multiple species but will run one by one</th>
<th>4: one species at a time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1 &amp; 4</td>
</tr>
<tr>
<td>INPUT DATA</td>
<td>SDtool</td>
<td>BioSim Tool</td>
<td>Fishpi4WKBIOPTIM</td>
<td>SimPop</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>-------------</td>
<td>------------------</td>
<td>--------</td>
</tr>
<tr>
<td>RDB</td>
<td>RDB (CS, CA, TR, HH, HL, CL)</td>
<td>RDB (CA, HH)</td>
<td>RDB (HL, HH, SL, TR)</td>
<td>RDB (CS, CA, TR, HH, HL, CL)</td>
</tr>
<tr>
<td>Other format</td>
<td>Cost objects (CS and CL)</td>
<td>N/A</td>
<td>FishPi data call</td>
<td>No</td>
</tr>
</tbody>
</table>

<p>| OUTPUTS | Quality indicators (eg EMD) | EMD, CV | MWCV, EMD, Number of size/age/sex classes sampled, Number of modes. | RSE, MWCV (raw outputs allow to calculate user-defined outputs) | EMD, CV, MWCV | see section 2.3 of WKBIOPTIM3 | see section 2.1 of WKBIOPTIM1 | L1-distance (ADV) |</p>
<table>
<thead>
<tr>
<th>Example Question</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sdtool</strong></td>
</tr>
<tr>
<td><strong>BioSim Toom</strong></td>
</tr>
<tr>
<td><strong>Fishpi4WKBIOPTIM</strong></td>
</tr>
<tr>
<td><strong>SimPop</strong></td>
</tr>
<tr>
<td><strong>SampleOptim</strong></td>
</tr>
<tr>
<td><strong>SampleLevelOptim</strong></td>
</tr>
<tr>
<td><strong>SampleReferenceLevel</strong></td>
</tr>
</tbody>
</table>
5 Case studies

5.1 SampleReferenceLevel Tool: Application in three case studies

5.1.1 Plaice in subarea 27.4.b, Quarter 4 (2018): German fleet length sampling data

1. Germany samples plaice, given its status as a G1 species. So, it can be often sampled in excess due to conservative planning. The length distribution of the original dataset is presented in Figure 5.1.1.1.

Figure 5.1.1.1 - Histogram of length frequency distribution (LFD), raised whole catch, \( \Delta = 1 \text{ cm} \).

Figure 5.1.1.2 - Histogram of length frequency distribution (LFD), raised whole catch, under larger bin width, \( \Delta = 3 \text{ cm} \).
2. Robust critical points:
   - Modes: 7, 31
   - Antimodes: 15

3. Reference subsample on set $l'$, constructed under parameters values $\Delta = 3$ cm, $\theta = 0.9$, $\varepsilon = 1$, $l' = [0; 40]$, $\gamma = 0.3 \, n_{l_j}$ where $n_{l_j}$ is a number of individuals in the length class $l_j$ in original sample, $l_j \in l'$.

![Figure 5.1.1.3](image)

Figure 5.1.1.3 – Length frequency distributions (LFD), raised whole catch, for the original sample (left) and the reference subsample (right), for length classes $\leq 40$ cm.

This analysis showed that we can delete 954 individuals without changing distributional shape, from length classes less or equal 40 cm.
4. Threshold computation: \( \text{ADV} = 0.29 \)

![Empirical cumulative distribution functions of the original sample (pink) and the reference subsample (green).]

5. Effort optimization: Trips sampling reducing.

**Table 5.1.1.1** - Trips participating in plaice length sampling in the given area/quarter/year. In brackets the numbers from length classes less or equal 40 cm are given.

<table>
<thead>
<tr>
<th>Trip number</th>
<th>Number measured individuals (raised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEU20180683479</td>
<td>616 (616)</td>
</tr>
<tr>
<td>DEU20186830700</td>
<td>26178 (25609)</td>
</tr>
<tr>
<td>DEU20186832100</td>
<td>161 (161)</td>
</tr>
<tr>
<td>DEU20187835788</td>
<td>2126 (1960)</td>
</tr>
</tbody>
</table>

Analysis based on assumed criteria demonstrates that the trip DEU20187835788 can be excluded from sampling process (Figure 5.1.1.5).
The real-world subsample meets the conditions from Definition 2 and the distance between its CDF function and the CDF of original data less than ADV.

5.1.2 Haddock in Division 27.7.e, Quarter 1 (2015): English fleet length sampling data

1. Original data set is presented in Figure 5.1.2.1

Figure 5.1.2.1 - Histogram of length frequency distribution (LFD), raised whole catch, $\Delta = 1 \text{ cm}$.
The length distribution under larger bin width ($\Delta = 5 \text{ cm}$) is presented in Figure 5.1.2.2.

![Histogram of length frequency distribution (LFD), raised whole catch, $\Delta = 5 \text{ cm}$.](image)

Figure 5.1.2.2 - Histogram of length frequency distribution (LFD), raised whole catch, $\Delta = 5 \text{ cm}$.

2. Robust critical points:
   - Modes: 33, 49, 62
   - Antimodes: 42, 59

3. Reference subsample on set $l'_{ij}$, constructed under parameters values $\Delta = 5 \text{ cm}$, $\theta = 0.7$, $\varepsilon = 0$, $l' = [30; 70]$, $\gamma = 0.3 n_{ij}$, where $n_{ij}$ is a number of individuals in the length class $l_j$ in original sample, $l_j \in l'$ (Figure 5.1.2.3).
Thus saving 5676 measurements will not change a considered LFD shape.

4. Threshold computation: \( \text{ADV} = 2.82 \)
5. **Effort optimization: Fleet sampling reducing.**

Table 5.1.2.1 - Fleets participating in haddock length sampling in the given area/quarter/year. In brackets the numbers from length classes between 30 and 70 cm are given.

<table>
<thead>
<tr>
<th>FAC_EC_lvl6</th>
<th>Number measured individuals (raised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNS_DEF_120-219_0_0</td>
<td>3 (3)</td>
</tr>
<tr>
<td>OTB_DEF_70-99_0_0</td>
<td>273 (273)</td>
</tr>
<tr>
<td>OTB_MOL_70-99_0_0</td>
<td>1 (1)</td>
</tr>
<tr>
<td>OTT_DEF_70-99_0_0</td>
<td>2 (2)</td>
</tr>
<tr>
<td>PTB_DEF_100-119_0_0</td>
<td>18551 (17884)</td>
</tr>
<tr>
<td>PTB_DEF_70-99_0_0</td>
<td>111 (111)</td>
</tr>
<tr>
<td>TBB_DEF_70-99_0_0</td>
<td>447 (446)</td>
</tr>
<tr>
<td>TBB_MOL_0_0_0</td>
<td>659 (655)</td>
</tr>
</tbody>
</table>

Considered length sample may be possibly reduced to two métiers, PTB_DEF_100-119_0_0 and TBB_DEF_70-99_0_0 (see Figure 5.1.2.5).

![Figure 5.1.2.5 – Empirical cumulative distribution functions of the original sample (purple) and the real-world subsample (red).](image)

**5.1.3 Plaice in Division 27.7.d, all quarters (2018): Belgian fleet length sampling data**

ILVO is collecting fishery-dependent data for the beam trawl fleet targeting demersal species (TBB_DEF). Catch information is obtained through ‘At sea sampling’. When an observer is onboard a vessel, every second haul is sampled so sampling takes place around the clock to reflect typical working conditions. The observer sorts all species of commercial importance and
determines the total weight by species and catch category. For a selected set of species a.o. plaice
(*Pleuronectes platessa*), the observer also takes length measurements.

ILVO is investing in optimizing the design of the at sea sampling programme. The ultimate goal
is to design a new statistically sound sampling programme that will follow the guidelines for
good practice provided by ICES and the European Commission (making the most efficient use
of sampling resources in order to collect unbiased and precise catch data) and implement this
new programme in practice.

For this case study, information of plaice (*Pleuronectes platessa*) caught with TBB_DEF_70-99_0_0
in 2018 (all quarters) in the Eastern English Channel was used. The data set covers 13 sampled
trips and 195 sampled hauls with length frequency information (47071 length measurements). In
this case study, we will test if the amount of sampled hauls within a trip can be reduced without
losing important information.

1. Original data set is presented in Figure 5.1.3.1

![Figure 5.1.3.1 - Histogram of length frequency distribution (LFD), raised whole catch, \(\Delta = 1\) cm.](image)
2. Robust critical points:
   - Modes: 24
   - Antimodes: 0

This is a special case of “one big bump”. In Belgian length sampling program, the results from all quarters are combined for the future analysis, and that defines a given LFD form. In this case we set a single antimode to 0. Obviously, the central length classes are oversampled, so we try to see how can be sampling effort reduced, even if we change an amplitude value slightly (0.95 instead of 1).

3. Reference subsample for all length classes, constructed under parameters values $\Delta = 2$ cm, $\theta = 0.95$, $\varepsilon = 0$, $\gamma = 0.3 \, n_{ij}$, where $n_{ij}$ is a number of individuals in the length class $l_j$ in original sample (Figure 5.1.3.3).
Thus saving $23\,846 = 179\,914 - 156\,068$ measurements will not change a considered LFD shape.

4. Threshold computation: $\text{ADV} = 0.61$

5. Effort optimization: Hauls sampling reducing.

We will eliminate a number of sampled hauls from one trip, e.g. NDGP18_S_PTCT01. So, we omit each second haul of that trip (see Figure 5.1.3.5).

The conclusion is that sampling of these hauls might be reduced.
So, the approach establishes quantitative objectives of length sampling not directly in the form of sample size determination, but in the form of selection of “information-sufficient”, adequate length distribution. Justification of distributional adequacy is however not a trivial task and can involve some kind of subjectivity (actually, as well as optimal sample size determination). Focusing on practical reasons and aspects will minimize a subjectivity. Indeed, if we define our practical goals for specific species as we made in example above, i.e. “subsample approximately close to the entire national sampling data at trip level”, the concept of the adequate distributional shape will be based on reference subsample produced from aggregated results of national sampling process under pre-specified parameters. However, practical goals might be guided by regional sampling strategy principles, where the requirements for LFD precision is at the regional métier level. Then, to achieve a certain level of precision established by this reference subsamples, some redesigns of sampling models might be recommended.

5.2 Case study: Greece Purse seine (WKBIOPIM2 Multi-level analysis & SDTool)

5.2.1 WKBIOPIM2 Multi-level analysis

This 1st part case study relies on the tool developed during WKBIOPIM2 for Multi-level analysis of sampling effort needed to produce estimates of catch numbers for stock assessment (Section 2.5).

A multi-level analysis was performed on the purse seine Greek data available on the area GSA22 (Aegean Sea) for the year 2014. The Greek fleet of purse seiners consist of 258 vessels (~1.6% of the total fleet, 1/7/2014 data) and ~215 of them operate in the GSA22, targeting pelagic species, mainly anchovies (Engraulis encrasicolus) and sardines (Sardina pilchardus). Fishing for purse seiners in Greece is restricted in winter, between 15 December and 28 February.
5.2.1.1 Sampling design
For purse seiners, landings are collected using the national fishery declaration system for all the vessels trips (Total Landings for 2014 in Figure 5.2.2.1). For biological data, the sampling scheme is spatially, temporary (quarterly), and per métier stratified. The Primary Sampling Unit (PSU) is the fishing trip (in the analyses only the on board sampling trips were used), which is randomly selected (with replacement) within each stratum. Within a fishing trip, all hauls are recorded. The catch fractions that are covered are Landings (on-shore and on-board sampling) and Discards (on board sampling). Inside each catch fragment, in practice, all species are recorded.

5.2.1.2 Data
Landings per quarter (10 species with the highest annual landings) are presented in Figure 5.2.1.2.1.

In Figure 5.2.1.2.2 the number of sampled trips and the total landings per species are presented. The dataset contained 13 species. To select the most important species for the fishery in order to proceed to the analyses, we set a threshold for the number of trips where each species was recorded (> 100) as well as for the total landings per year (> 100 tonnes). Based on these criteria, three species were selected, *E. encrasicolus*, *S. pilchardus* and Atlantic chub mackerel (*Scomber colias*). Anchovy and sardine, as we mentioned above, are the main target for purse seines, while Atlantic chub mackerel is an important secondary target species for this fishery.
Figure 5.2.1.2.2 - Total landings and N of trips that recorded each species. With red colour the species selected for further analysis.

5.2.1.3 Exploratory data analysis

Since only one métier in one region has been selected for the simulations (the lowest levels of the technical and spatial stratification hierarchy), only the temporal stratification could be considered for the analyses. Figure 5.2.1.3.1 shows the selected species represented by the total landings and the number of trips sampled per quarter. The main pressure on the stocks is detected during the second and third quarter of the year, while in the first and fourth quarter, which coincide with the restriction period, landings are significantly lower. From the second to the fourth quarter, the three species seem to be almost omnipresent in the sampling trips catch. It must be noted also, that the sampling seems to be somewhat unbalanced; based on the volume of landings, it seems that lower sampling effort should be allocated to the fourth quarter in respect to the second and third.
Figures 5.2.1.3.2-5.2.3.4 present the length frequency distribution for the three species raised to the population (per quarter), along with the corresponding CV values for every length class. As it was expected, Number of Individuals and CV are, more or less, inversely proportional, which means that in abundant length classes CV is low and vice versa. More interestingly, the length frequency distribution (LFD) seems to be quarter-dependent (e.g. for *E. encrasicolus*, the pick of the LFD is at Total Length (TL) equal to 120 mm in quarters 2 and 3 and at TL 110 mm for the fourth quarter). For *E. encrasicolus* and *S. pilchardus*, the seasonal LFDs, despite their differences, have a high degree of overlap. On the contrary, *S. colias* LFDs exhibit quite different patterns between different quarters: the pick in 2nd quarter is at TL 250 mm, in 3rd quarter at 230 mm and in fourth at 170 mm, with a low degree of overlap. These seasonal fluctuations should be attributed to the biology of the species, to the effect of the restriction period, to environmental seasonal fluctuations and/or to random events.
Figure 5.2.1.3.2 - Length structure estimates for *E. encrasicolus* raised to the population, along with the corresponding CV values.

Figure 5.2.1.3.3 - Length structure estimates for *S. pilchardus* to the population, along with the corresponding CV values.
5.2.1.4 Need for optimization
In order to test the effects of down-sampling the number of trips, bootstrapping simulations were run during which a gradually increasing proportion (0.1 to 1, step = 0.1) of the total number of trips were sampled. For each proportion, a corresponding number of simulations were run (N = 50). For each simulation, LFD and weighted coefficient of variation (WCV) were calculated. Our goal was to explore if and how down-sampling would affect the quality of LFD data for each species in terms of the corresponding CV. First quarter was omitted from the simulations due to low sampling effort (N = 3 trips).

5.2.1.5 Results
The total number of trips per year (2014) for GSA22 was 147. For *E. encrasicolus* and *S. pilchardus*, the overall shape of the seasonal LFDs was not affected significantly by down-sampling (Figures 5.2.1.5.1 and 5.2.1.5.3). As it was expected, the CV for the most abundant length classes didn’t suffer severe changes due to down-sampling, at least at rates equal to 0.5 to 0.75 times total number of trips (yellowish dots in Figures 5.2.1.5.2 and 5.2.1.5.4). However, down-sampling seem to have a more significant effect on the lower and upper bound classes.
Figure 5.2.1.5.1 - Bootstrap simulation outcomes on Length Frequency Distribution changes for *S. pilchardus*. The number of sampled individuals per length class as well as the fitted lines for the LFD corresponding to each down-sampling proportion (scale of blue) are depicted.

Figure 5.2.1.5.2 - Bootstrap simulation outcomes on CV changes for *S. pilchardus*. The corresponding CV values per length class for each simulation is depicted. The coloured scale indicates the proportion of down-sampling in respect to the total number of trips.
Figure 5.2.1.5.3 - Bootstrap simulation outcomes on Length Frequency Distribution changes for *E. encrasicolus*. The number of sampled individuals per length class as well as the fitted lines for the LFD corresponding to each down-sampling proportion (scale of blue) are depicted.

On the contrary, both the shape and CV values for *S. colias* seem to be very sensitive to down-sampling (Figures 5.2.1.5.5 and 5.2.1.5.6). The shape of the LFD changes between the various proportions of down-sampling and CV values fluctuate over a wide range of values. Moreover, the risk of completely losing some lower and upper band length classes due to down-sampling, is high for this species.
Figure 5.2.1.5.5 - Bootstrap simulation outcomes on Length Frequency Distribution changes for *S. colias*. The number of sampled individuals per length class as well as the fitted lines for the LFD corresponding to each down-sampling proportion (scale of blue) are depicted.

Figure 5.2.1.5.6 - Bootstrap simulation outcomes on CV changes for *S. colias*. The corresponding CV values per length class for each simulation is depicted. The coloured scale indicates the proportion of down-sampling in respect to the total number of trips.

Figure 5.2.1.5.7 presents the evolution of the MWCV of the length distribution in relation to the number of trips used to compute the estimates. WCV is a metric that can quantify the overall effect of down-sampling on the LFD. In consistency with the previous observations, WCV for the two main targets of the fishery (anchovy and sardine) remain quite stable as the simulated
sampling effort decreases, exhibiting low to medium fluctuations even at down-sizing proportions equal to 0.5. On the other hand, the risk of losing information for *S. colias*, is quite high, even at the lowest values of down-sampling (0.9 times sampling effort).

![Figure 5.2.1.5.7 - Bootstrap simulation outcomes on weighted CV changes for the three species in respect to the down-sampling proportion. The fitted line as well as confidence intervals for each species is also depicted.](image)

### 5.2.1.6 Conclusions

The outcomes of our simulations highlight the complex nature of the sampling process within the Data Collection Framework premises. The use of a bootstrapping simulation technique to improve the sampling plan (in terms of number of trips sampled), revealed that for the main targeted species anchovy and sardine (landed in higher quantity and sampled accordingly), there is wide room for improvement (down-sampling wouldn’t have serious consequences on the accuracy of the data). On the other hand, the consequences on a secondary target species for the fishery, *S. colias*, seem to be more important, since a slight modification on the number of sampled trips can decrease dramatically the quality of the length distribution estimates and more importantly can lead to the absence of lower and upper bound length classes in the samples.

Simulations on this case study was performed only on landings data and the criterion to assess the loss of information due to down-sampling was the changes on LFD in terms of the corresponding CV. However, on-board sampling, apart from collecting data to construct reliable LFDs, has a variety of goals, such as the estimation of total discards, calculation of the maturity ogives, sex ratio, age-length keys as well as collecting data on rare Protected, Endangered and Threatened species. Therefore, any attempt to reduce sampling effort should consider the effects on a wide range of important analyses for stock assessment.
5.2.2 SDTool

5.2.2.1 Fishery or data description
Commercial data on length from onboard sampling of anchovy (*E. engrasiculus*) and sardine (*S. pilchardus*) in the Aegean (GSA 22) and Eastern Ionian (GSA 20) Sea for the years 2014 – 2016 was used to derive the optimal number of trips for annual sampling for each GSA. Anchovy and sardine in Greece is almost entirely caught by purse seiners. The Greek fleet of purseiners consists of around 260 vessels, while the vast majority of them (~85%) operate in the Aegean Sea. There is a temporal closure of purse seine fishery in Greece from the 15 December to the 28 February.

5.2.2.2 Sampling programme
The data available were collected by quarter and by gear although for anchovy only purseiners were considered adequate for the analysis. The tested hypothesis is 12 trips as minimum per year for both areas in order to take into consideration the monthly variability of the length frequency distribution of the two species.

5.2.2.3 Need for optimization (what do you want to test)
The aim of this case study is to test the effect of down-sampling the number of trips hold annually and try to optimize both for anchovy and sardine, as these two species are the most abundant in purse seine fisheries. SDTool’s output is a range to derive optimal number of trips.
5.2.2.4 Interpretation of results

Figure 5.2.2.4.1 - Exploratory graphs of CV against number of trips for anchovy in GSA 20 (above) and GSA 22 (below).
Figure 5.2.2.4.2 - Exploratory graphs of CV against number of trips for sardine in GSA 20 (above) and GSA 22 (below).

Exploratory graphs of Coefficient of Variation against the number of trips for anchovy suggest around 30 trips per year for GSA 20 and slightly above 35 for GSA 22 in order to achieve the threshold of 20% CV.
In the figures above, the density of the CV used by the method to infer the solutions (optimal number of trips), is presented. The red lines indicate all the solutions inferred as “optimal” sampling size and for each line the number of trips and the min-max value of the CV are reported. While a minimum number of trips was set to 12, a CV of around 0.2 is considered acceptable. Hence, although in GSA 22, 12 number of trips is among the optimal solutions is rejected due to very high CV. For GSA 20 the optimal number of trips is around 20 trips per year and for GSA 22 a number around 40 trips per year results in an acceptable CV.

The graphs above indicate an optimal number of trips of around 40 trips per year for sardine in GSA 20 while for sardine in GSA 22 seems that the optimal solution for the number of trips is above the 35 trips per year.
Figure 5.2.4.5 - CV versus the number of trips for anchovy in GSA 20 (above) and GSA 22 (below). Vertical blue lines indicate the first 3 solutions in terms of number of trips (greater than the minimum accepted solution indicated by the user as acceptable) inferred via the optimization process.

Solutions for the range of optimal number of trips indicate that the current sampling is quite higher in terms of numbers of sampling by year for the anchovy both areas while after the 60 trips per year the CV reaches a plateau of 0.1.
Figure 5.2.2.4.6 - CV versus the number of trips for sardine in GSA 20 (above) and GSA 22 (below). Vertical blue lines indicate the first 3 solutions in terms of number of trips (greater than the minimum accepted solution indicated by the user as acceptable) inferred via the optimization process.

The same picture as for the anchovy appears for the sardine, with the solutions derived from the optimization process indicating an oversampling for both years and areas.
Table 5.2.2.4.1 Optimal number of sampling trips and their corresponding min and max CVs for anchovy in GSA 20 and GSA 22.

<table>
<thead>
<tr>
<th>Area</th>
<th>solutions</th>
<th>maxCV</th>
<th>minCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSA20</td>
<td>19</td>
<td>0.38</td>
<td>0.1</td>
</tr>
<tr>
<td>GSA20</td>
<td>38</td>
<td>0.3</td>
<td>0.12</td>
</tr>
<tr>
<td>GSA20</td>
<td>54</td>
<td>0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>GSA20</td>
<td>&gt;54</td>
<td>0.16</td>
<td>0</td>
</tr>
<tr>
<td>GSA22</td>
<td>12</td>
<td>0.46</td>
<td>0.26</td>
</tr>
<tr>
<td>GSA22</td>
<td>34</td>
<td>0.41</td>
<td>0.13</td>
</tr>
<tr>
<td>GSA22</td>
<td>41</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>GSA22</td>
<td>&gt;41</td>
<td>0.26</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 5.2.2.4.2 - Optimal number of sampling trips and their corresponding min and max CVs for sardine in GSA 20 and GSA 22.

<table>
<thead>
<tr>
<th>Area</th>
<th>solutions</th>
<th>maxCV</th>
<th>minCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSA20</td>
<td>28</td>
<td>0.39</td>
<td>0.1</td>
</tr>
<tr>
<td>GSA20</td>
<td>39</td>
<td>0.19</td>
<td>0.1</td>
</tr>
<tr>
<td>GSA20</td>
<td>57</td>
<td>0.16</td>
<td>0.06</td>
</tr>
<tr>
<td>GSA20</td>
<td>&gt;57</td>
<td>0.11</td>
<td>0</td>
</tr>
<tr>
<td>GSA22</td>
<td>34</td>
<td>0.36</td>
<td>0.13</td>
</tr>
<tr>
<td>GSA22</td>
<td>&gt;34</td>
<td>0.2</td>
<td>0</td>
</tr>
</tbody>
</table>

Small pelagic fisheries in Greece is characterized of variability of species and especially for anchovy and sardine, the fishing grounds are not always distinctively separated. Given that and that the fact that the target of the fisheries is almost never single species a combine solution of the above tables should be conducted.
Figure 5.2.4.7 - Recycling rate of the samples against the number of trips by year for anchovy in GSA 20 (above) and GSA 22 (below). Vertical blue lines are referred to the first 3 solutions in terms of number of trips (greater than the minimum accepted solution indicated by the user as acceptable) inferred via the method. Red line corresponded to an arbitrary critical threshold of 50%.
Figure 5.2.4.8 - Recycling rate of the samples against the number of trips by year for sardine in GSA 20 and GSA 22. Vertical blue lines are referred to the first 3 solutions in terms of number of trips (greater than the minimum accepted solution indicated by the user as acceptable) inferred via the method. Red line corresponded to an arbitrary critical threshold of 50%.
5.2.2.5 **Next steps**
Include in the analysis small pelagic species that are targeted from the Greece purse seiners and have a larger range of length distribution. Apply also BioSim Tool in both cases for further investigation.

5.2.3 **Comparison between the two approaches**

Two different tools assessing sampling intensity, WKBIOPTIM2 Multi-level analysis and SDtool, were applied on the same database (Greek sampling on Purse Seine in GSA 22, in 2014). This fact provides the opportunity to assess and compare the performance of the two methods as well as their explanatory power on a common basis. The outcomes of the two methods are quite comparable, indicating oversampling when anchovy and sardine are the only two stocks considered. To a certain extent, this fact serves as a criterion for confirming the explanatory power, as well as the accuracy of the results of the two methods. SDtools’ advantage is that it is able to provide a specific suggestion as far as how many samples are the optimal for the case study, as well as the fact that it is able to simulate different scenarios of sampling intensity, while Multi-level analysis is based on the visual inspection of graphs. On the other hand, the fact that multi-level analysis can simulate information from many species simultaneously, could be proven particularly useful, since it enables the detection of a wider range of possible reduction rates on the sampling effort. Since the two tools have comparable power in terms of quality of results, the choice between the two is based on the needs of the potential user.

5.3 **M. barbatus in Southern Adriatic Sea (application of SampleOptim and BioSim Tool)**

5.3.1 **Fishery or data description**
Commercial data on length, age and sex of red mullet (*M. barbatus*) in GSA 18 (Southern Adriatic Sea) for the years 2014-2017 was used to derive the optimal number of otoliths by length class to be read according to an annual and a quarterly sampling.

5.3.2 **Sampling programme**
The data available were collected by quarter, sex and year (5 otoliths by sex and 0.5 cm length class). The common tested hypotheses for both applications are the yearly and the quarterly sampling of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 otoliths by length class. The analyses were carried out separately for each sex.

5.3.3 **Need for optimization (what do you want to test)**
The aim of this case study is to test the SampleOptim and the BioSim Tool to verify the consistency of their results on the same dataset and highlight the peculiarities of each tool. Indeed, if on one hand the BioSim Tool allows to have an idea of the impact of changing the number of age readings on sampling precision, also working at sample level (by trip) to derive synthetic results, on the other hand the SampleOptim informs concerning the impact on the stability of the von Bertalanffy growth parameters and on the maturity ogive parameters (L25, L50 and L75) derived by those age readings, allowing to work at quarter, year and semester level and also considering a stratification on sexes (by defining the sex ratio) and at sampling Port level. The 2
tools can be considered as complementary, in case of considering only the temporal level stratification (quarter, semester and year), as in this case study.

5.3.4 Interpretation of results

SampleOptim
The results showed that for a quarterly age sampling, to collect about 3-4 otoliths would allow, for both sexes, to satisfactory stabilize the estimates of $L_{\infty}$, $K$ and $t_0$ (Figure 5.3.4.1 and Figure 5.3.4.2), while for a yearly age sampling 7-8 otoliths would be sufficient (Figure 5.3.4.4 and Figure 5.3.4.5). This numbers by length class would corresponds to an acceptable small Mean absolute percentage error (MAPE) (Figure 5.3.4.3 and Figure 5.3.4.6).

Figure 5.3.4.1 – Impact on the variability of von Bertalanffy coefficients in females according to different sampling scenarios, quarterly sampling.
Figure 5.3.4.2 – Impact on the variability of von Bertalanffy coefficients in males according to different sampling scenarios, quarterly sampling.

Figure 5.3.4.3 – Mean absolute percentage error (MAPE) of females (left) and males (right) in the hypothesis of quarterly sampling.
Figure 5.3.4.4 – Impact on the variability of von Bertalanffy coefficients in females according to different sampling scenarios, annual sampling.
Figure 5.3.4.5 – Impact on the variability of von Bertalanffy coefficients in males according to different sampling scenarios, annual sampling.

Figure 5.3.4.6 – Mean absolute percentage error (MAPE) of females (left) and males (right) in the hypothesis of annual sampling.
BioSim Tool
The results showed that in the hypothesis of a quarterly sampling, for both sexes the precision is more stable from 3–4 otoliths by length class, while for the annual sampling, from 7–8, approaching the current levels of CV.

Both hypotheses showed that a greater number of otoliths would be needed if the length classes are pooled across sexes, having obtained in both cases a higher CV compared to the one of the sexed individuals.

Moreover, it is also quite evident that a sampling by quarter would allow to obtain a more efficient age sampling in terms of precision, respect to the annual sampling.

![Figure 5.3.4.7 – ALK CV associated to each age sampling scenario in the hypothesis of a quarterly and an annual sampling.](image)

5.3.5 Next steps
For the dataset used for red mullet in GSA 18, the results obtained by the 2 tools seem sufficiently consistent. The next steps could be:

1. Test the 2 tools on different case studies;
2. A vignette containing an example of communication between the 2 tools (as in this case study) could help the user in the interpretation of the results provided by the 2 tools in a complementary way;
3. The development of an integrated tool, incorporating the features of BioSim (only the script related to age) and SampleOptim could be developed in the future, in order to combine the advantages of both.

5.4 *M. barbatus* in Southern Adriatic Sea (application of FishPi4WKBIOPTIM and SDTool)
The aim of this case study was to verify the consistency of the outputted results of FishPi4WKBIOPTIM and SDTool with respect to identifying optimal sampling schemes.
5.4.1 Data

2016–2017 commercial data on length of red mullet (M. barbatus) in GSA 18 (Southern Adriatic Sea) were used to assess the optimal number of trips to sample according to an annual or a quarterly sampling scheme. The available data were collected by quarter, and, for the 2 selected years, all quarters were sampled.

5.4.2 Approach

Two sampling schemes of interest were tested using FishPi4WKBIOPTIM and SDTool, each with a range of sampling effort, i.e. 2 to 6 trips per quarter or 8 to 24 trips per year. Both tools were further used to explore optimal solutions in terms of number of trips to sample for each sampling scheme.

FishPi4WKBIOPTIM was set to sampling with replacement, to match the SDTool settings. SDTool returns a set of solutions to derive an optimal range of number of trips to sample. FishPi4WKBIOPTIM was used to explore the range of scenarios mentioned above as well as to evaluate the impact on the estimated mean length of the optimal number of trips estimated by the SDTool. The range of scenarios were also explored using SDTool to derive the impact on the CV of alternative sampling strategies.

5.4.3 Results

SDTool

SDTool returns the first 4 solutions in terms of number of trips to be sampled to optimize the precision of the sampling (CV). The results showed that the current sampling was in line with the first solution provided by SDTool for quarters 1, 2 and 4, in quarter 3 the variability on the LFDs increased because of the recruitment event (Figure 5.4.4.1, Table 5.4.4.1). For the sampling on an annual time step, the current annual number of trips was closer to the second solution of the optimal range, showing less dispersion around the mean CV (Figure 5.4.4.2, Table 5.4.4.2).
Figure 5.4.4.1 – CV obtained by the resampling procedure. The black line is the interpolated curve and the vertical lines correspond to the first solutions detected to derive the optimal range for the number of trips by quarter assuming a quarterly sampling from SDTool. From top left to bottom right quarter 1 to quarter 4.
Figure 5.4.4.2 – CV obtained by the resampling procedure. The black line is the interpolated curve and the vertical lines correspond to the first solutions detected solutions to derive the optimal range for the number of trips by year assuming a yearly sampling from SDTool.

For the quarterly sampling only the first solution of each quarter was considered in the next phase, while for the annual sampling, both the first and the second solutions will be considered in the next analyses.

Table 5.4.4.1 – First solution by type of sampling by SDTool by quarter (only the first solution by quarter) and by year (first two solutions).

<table>
<thead>
<tr>
<th>Sampling</th>
<th>Quarter</th>
<th>Metier</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>quarterly</td>
<td>1</td>
<td>OTB_DEF</td>
<td>8</td>
</tr>
<tr>
<td>quarterly</td>
<td>2</td>
<td>OTB_DEF</td>
<td>9</td>
</tr>
<tr>
<td>quarterly</td>
<td>3</td>
<td>OTB_DEF</td>
<td>18</td>
</tr>
<tr>
<td>quarterly</td>
<td>4</td>
<td>OTB_DEF</td>
<td>12</td>
</tr>
<tr>
<td>yearly</td>
<td>Year</td>
<td>OTB_DEF</td>
<td>19</td>
</tr>
<tr>
<td>yearly</td>
<td>Year</td>
<td>OTB_DEF</td>
<td>40</td>
</tr>
</tbody>
</table>

The exploration of alternative sampling designs was carried out according to the results obtained from the optimization:

1. Baseline: number of trips by quarter estimated as average of the number of trips by quarter of the 2 available years (Q1 = 9, Q2 = 8; Q3 = 9; Q4 = 14, for a total of 40 trips);
2. First_sol_YEAR: resampling by year, according to the first solution by year (19 trips);
3. Second_sol_YEAR: resampling by year, according to the second solution by year (40 trips);
4. Quarter_opt: resampling by quarter, according to the optimal solution by quarter.
The SDTool results implied that the current annual number of trips sampled was in line with the optimal value derived from SDTool; it is also worth noting that a reduction of effort by half increased the CV by only 3%. The sampling scheme (by quarter or by year) did not seem to significantly influence the CV, the corresponding number of individuals needed for length measurements or the raised LFDs (Table 5.4.4.2, Figure 5.4.4.3).

**Table 5.4.4.2 – Performances in terms of precision, number of trips and number of individuals for which the length has to be collected by scenario carried out with SDTool.**

<table>
<thead>
<tr>
<th>species</th>
<th>GSA</th>
<th>Metier</th>
<th>scenario</th>
<th>CV</th>
<th>samp_size</th>
<th>no_indiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mullus barbatus</td>
<td>GSA18_ITA</td>
<td>OTB_DEF</td>
<td>Baseline</td>
<td>24.01</td>
<td>40</td>
<td>11770</td>
</tr>
<tr>
<td>Mullus barbatus</td>
<td>GSA18_ITA</td>
<td>OTB_DEF</td>
<td>First_sol_YEAR</td>
<td>27.06</td>
<td>19</td>
<td>5471</td>
</tr>
<tr>
<td>Mullus barbatus</td>
<td>GSA18_ITA</td>
<td>OTB_DEF</td>
<td>Second_sol_YEAR</td>
<td>22.54</td>
<td>40</td>
<td>11655</td>
</tr>
<tr>
<td>Mullus barbatus</td>
<td>GSA18_ITA</td>
<td>OTB_DEF</td>
<td>Quarter_opt</td>
<td>22.91</td>
<td>47</td>
<td>13521</td>
</tr>
</tbody>
</table>

**Figure 5.4.4.3 - Performance of the explored sampling scenarios in terms of CV (the value is reflected by the magnitude of the points), defined in terms of number of trips (x-axis) and number of measured individuals (y-axis) (left). On the right the comparing of the raised LFD derived from the 4 explored scenarios is shown. The dotted lines represent the current sampling (baseline).**

**FishPi4WKBIOPTIM**

In FishPi4WKBIOPTIM, a quarterly sampling scheme lead to a lower RSE across the different scenarios explored (Figure 5.4.4.4 and Figure 5.4.4.6) compared to a yearly approach for this stock. Despite this, both yearly and quarterly sampling schemes showed similar results in terms of estimated mean length in 2016. In 2017 mean length resulting from the yearly sampling was closer to the estimated mean length of the population, i.e. original dataset it was sampled from (Figure 5.4.4.5 and Figure 5.4.4.7). The same results were obtained if the sampling was set without replacement. In both cases 100 simulations were run.
Figure 5.4.4.4 - Results of the simulation in terms of RSE for all the sample size scenarios, by year (left) and aggregated (right) for a quarterly sampling.

Figure 5.4.4.5 - Results of the optimization in terms of estimated mean length (for the scenario with 6 trips per quarter) for a quarterly sampling.
The exploration of the scenarios suggested by the SDTool in FishPi4WKBIOPTIM showed that the comparison between the estimated mean length in the current sampling and in the tested scenarios is more consistent using the number of trips of the second solution (40 trips by year) by year and of the number of trips by quarter (first solution) (Figure 5.4.4.8 and Figure 5.4.4.9).
Results of the optimization in terms of estimated mean length for the scenario with the number of trips indicated by the SDTool by quarter.

SDTool and FishPi4WKBIOPTIM are consistent in that the sampling scheme (by quarter and by year) does not appear to influence the deviation of the estimated mean length from the mean of the full dataset.

### 5.4.4 Next steps

1. Test the 2 tools on different case studies;
2. A vignette containing an example of communication between the 2 tools (as in this case study) could help the user in the interpretation of the results provided by the 2 tools in a complementary way.
3. Test with the latest version of FishPi4WKBIOPTIM which provides actual CVs and length frequencies rather than just mean length (options developed after the workshop).
5.5  Sandeel in The greater North Sea (application of FishPi4WKBIOPTIM and SDTool)

5.5.1  5.5.1 Fishery or data description

*Ammodytes* are small demersal fish landed unsorted, being pumped directly into the factories for reduction. The greater North Sea covers seven different *Ammodytes* stocks, see Figure 5.5.1.1 and the Danish sandeel fishery lands *Ammodytes* from five different stocks see Table 5.5.1.1. 99.8% of the landings are caught with OTB.DEF.<16_0_0 and the fishing season is very limited, mainly taking part in the second quarter, see Figure 5.5.1.1. Further, the fishery is nearly solely landing *Ammodytes*, in 2018 these accounted for 98.9% of the landings from this fishery.

*Ammodytes* are fast growing species.

Denmark samples the sandeel fishery in two different sampling scheme – an at-sea self-sampling scheme and an on-shore sampling scheme.

**At-sea self-sampling:** in principle, all Danish fishers having a license for fishing sandeel should self-sample hauls for DTU Aqua. According to the present protocol number of sampled hauls depends on where the fishery is taking place - one sampled haul per day in area SA.1, SA.2, SA.3, and one sample from every haul in area SA.4 and SA.6. DTU Aqua subsamples the number of samples, so we end up with one per vessel, day and square. The present sampling scheme is set up to as a grid sampling aiming at satisfying the model for estimating catch at age and associated mean weight at age. This model runs with ½ month and ICES square at the lowest level.

**On-shore sampling:** DTU Aqua receives samples taken by the Danish fishery control during landing / pumping to the factories – in 2018; we worked up 17 of these samples. The selection of samples is based on the control’s protocol and we primarily collect these samples to have ‘fishers’ independent samples. These samples are not included in this study, since the design of that program is very different from the self-sampling program.

The at-sea self-sampling scheme yields many samples and puts a heavy workload on the laboratories, therefore it is a good candidate for optimization. DTU Aqua is in an ongoing process of optimizing the number of samples for this fishery and during WKBIOPTIM (2017) some of the WKBIOPTIM sample level tools were tested on the sampled length distributions. We are at the moment setting up at simulation framework where different scenarios are being judge on the resulting catch at age and mean weight at age matrix estimated with the present model. The main objective of this case study is to get familiar with the tools available from WKBIOPTIM, FishPi4WKBIOPTIM and SDTool 2.04. In particularly the quality indicators, but also simulation frameworks using design based estimation.
Figure 5.5.1.1 - Spatial distribution of the *Ammodytes* stocks (left). Danish landings of *Ammodytes* per month, 2018 (right)

Table 5.5.1.1 - Overview of Danish landings and sampling of *Ammodytes*, 2018

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter</th>
<th>Stock area</th>
<th>Danish landings (tons)</th>
<th>Sampled trips (self-sampling)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>1</td>
<td>SA.1</td>
<td>113,459</td>
<td>114</td>
</tr>
<tr>
<td>2018</td>
<td>1</td>
<td>SA.2</td>
<td>20,217</td>
<td>109</td>
</tr>
<tr>
<td>2018</td>
<td>1</td>
<td>SA.3</td>
<td>3,602</td>
<td>11</td>
</tr>
<tr>
<td>2018</td>
<td>1</td>
<td>SA.4</td>
<td>32,766</td>
<td>31</td>
</tr>
<tr>
<td>2018</td>
<td>1</td>
<td>SA.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2018</td>
<td>2</td>
<td>SA.1</td>
<td>6,376</td>
<td>1</td>
</tr>
<tr>
<td>2018</td>
<td>2</td>
<td>SA.2</td>
<td>285</td>
<td>0</td>
</tr>
<tr>
<td>2018</td>
<td>2</td>
<td>SA.3</td>
<td>581</td>
<td>1</td>
</tr>
<tr>
<td>2018</td>
<td>2</td>
<td>SA.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2018</td>
<td>2</td>
<td>SA.6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
5.5.2 Dataset description

Samples from the at-sea self-sampling program, 2018, described above. Only samples from the 2nd quarter have been used. Some overviews of the underlying samples are given in Figures 5.5.2.1–5.5.2.4.

Figure 5.5.2.1 - Length distribution by trip in area SA.1. Output from SDTool 2.04

Figure 5.5.2.2 – Length distribution by trip in area SA.2 output from SDTool 2.04.
5.5.3 Scenarios of simulations and results

STREAM SD TOOL 2.04
Strata: sandeel area, quarter and lev4 (same as lev6 in this case study)
nIter: 500
threshold_for_ReyclingRate: 0.5
Figure 5.5.3.1 – CV versus number of trips for SA.1

Figure 5.5.3.2 – CV versus number of trips for SA.2
Figure 5.5.3.3 – CV versus number of trips for SA.3

Figure 5.5.3.4 – CV versus number of trips for SA.4
FishPi2WP3_BioSims
Strata: sandeel area and quarter
Number of simulation: 500
With replacement: TRUE

Figure 5.5.3.5 – Relative standard error (RSE) versus number of trips (sampSize) per sandeel area
Figure 5.5.3.6 – Mean weighted CV (MWCV) versus number of trips (effort) per sandeel area.

Figure 5.5.3.7 – Distribution of mean length per re-sampled for a for a sample size of 20.
5.5.4 Conclusion

First, not all of the tools available in FishPi4WKBIOPTIM and SDTool have been tested, so the main outcome of the two tools in this case study is an evaluation of samples, fishing trips, needed to get an optimal CV around the mean length. Further, the FishPi4WKBIOPTIM tool has only been tested with the samples and not the whole population.

The results from the two tools are quite similar, but not equal. Both tools suggest that sandeel in area SA.1 and SA.2 are highly oversampled; SDTool suggest that little under 30 samples should be enough and FishPi4WKBIOPTIM suggest around 15 samples, both far below the present sampling effort of around 110 per area. Some of the difference may be because FishPi4WKBIOPTIM in this case study ignores the population, while SDTool raise the samples to population. Further, the underlying resampling and estimation methods are quite different, so an equal result would not be expected. Anyway, the results seem very reasonable with the stratification selected for this case study.

The two different quality measurements, RSE and MWCV from FishPi4WKBIOPTIM reveals quite different patterns, especially for SA.4, Figures 5.5.3.5 and 5.5.3.6. The reason why needs to be more carefully assessed.

For one of the stratum / domains, SA.2, the outcome of SDTool reveals two distinct groups, Figure 5.5.3.2. This indicate that it would be beneficial to split the two ‘groups’ into two separate strata. Having a closer look at the mean length over time reveals that this stratum has a quite low variation in mean length per week compared to samples from e.g. SA.1 and SA.4, Figure 5.5.3.9. It also reveals that the mean length is increasing over time, which may explain the ‘groups’.

Figure 5.5.3.8 – The overall length distribution for *Ammodytes*, original sample (red) and result from each resampling (blue) for a sample size of 20.
Whatever this pattern is driven by growth over time, spatial changes in the fishing pattern or a combination thereof is not apparent, but the variation could probably be handled by a finer timely and spatial stratification. This highlights the need for a more flexible specification of stratification, since the ones present available in the tools do not fit all stocks. Further, this pattern is not apparent from FishPi4WKBIOPTIM, which emphasizes that more visualizations would improve the evaluation of the outcomes and thereby the resulting designs.

To evaluate a sampling design based on the mean length seems a bit rough. For *Ammodites* the estimates of interest are the catch at age matrix, which requires a good estimation of the age distribution, and the mean weight at age. An evaluation based on the resulting length distribution seems promising and it would be great to compare the results from that with a resulting age distribution utilizing the tools developed under WKBIOPTIM for comparing distributions. For FishPi4WKBIOPTIM this is work in progress, see Figure 5.5.3.8. Further, none of the tools presently supports sampling design relevant for this sampling scheme e.g. fishing trip, haul, subsampling thereof and individuals for aging and weighting or even haul as the PSU. It would be beneficial if the sampling units and order thereof could be more freely specified.
5.6  *Mullus barbatus* in the Aegean Sea- Greece (Application of SampleOptimRDB, BioSim and LanAge_opt)

The purpose of this case study is to study the effect of reducing the number of age readings per length class of the Greek *Mullus barbatus* stock in the Aegean Sea (GSA 22). To this end, three tools are employed:

a)  SampleOptimRDB, considering optimization on age sampling based on the growth parameters, the age-length key (ALK), the mean age-at-length, the mean length-at-age,

b)  BioSim Tool, considering age sampling optimization based on the ALK

c)  LanAge_opt, considering age sampling optimization based on the age distribution of the landings

5.6.1  Fisheries description

Red mullet (*Mullus barbatus*) is an important demersal resource for the fisheries in the Aegean Sea (GSA 22). In this area, red mullet is a shared stock exploited by both Greek and Turkish trawlers, gillnetters and trammel netters and, to a lesser extent, by other commercial fisheries.

![Figure 5.6.1.1 – M. barbatus in GSA 22 (Greece). Landings by year, quarter and Fishing activity category European lvl 6.](image)

5.6.2  Sampling programme - Dataset available

The dataset provided contains Greek landings, effort and sampling data of *M. barbatus* in GSA 22 (Aegean Sea) for the years 2014–2016. The 2015 data were excluded from the analyses due to the incomplete DCF implementation in that year.
The biological data sampling for this stock is conducted annually in every GSA, without any further stratification on the spatial, temporal or technical (gear) scale. Quota sampling is conducted for age, whereby, according to the sampling design, at least 10 samples (otoliths) should be collected for each length class (1–2 otoliths per month), pooled across the sexes. For most of the length classes an average of 50 otoliths is collected, with the exception of the smaller (< 100 mm) and larger (> 220 mm) sizes, whereby the collection of individuals is hampered due to the relatively lower availability in the fishery, as shown in Table 5.6.2.1.

Table 5.6.2.1 - Number of age readings by year and length class for M. barbatus in GSA 22 (Greece)

<table>
<thead>
<tr>
<th>Length Class (mm)</th>
<th>2014</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>80</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>90</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>110</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td>120</td>
<td>49</td>
<td>75</td>
</tr>
<tr>
<td>130</td>
<td>61</td>
<td>85</td>
</tr>
<tr>
<td>140</td>
<td>65</td>
<td>89</td>
</tr>
<tr>
<td>150</td>
<td>58</td>
<td>81</td>
</tr>
<tr>
<td>160</td>
<td>51</td>
<td>82</td>
</tr>
<tr>
<td>170</td>
<td>49</td>
<td>73</td>
</tr>
<tr>
<td>180</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td>190</td>
<td>35</td>
<td>64</td>
</tr>
<tr>
<td>200</td>
<td>29</td>
<td>59</td>
</tr>
<tr>
<td>210</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>220</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>230</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>240</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>260</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
5.6.3 Optimization methods & results

The purpose of this case study is to investigate the potential for optimization in the age data collection without compromising the estimates for the ALK, the growth parameters and the Landings at age, for the Red Mullet stock in GSA 22 (Aegean Sea- Greece).

To this end, three tools are employed:

a) SampleOptimRDB, considering optimization on age sampling based on the growth parameters, the age-length key (ALK), the mean age-at-length, the mean length-at-age,
b) BioSim Tool, considering age sampling optimization based on the ALK

c) LanAge_opt, considering age sampling optimization based on the age distribution of the landings

5.6.3.1 SampleOptim for the Red Mullet stock in GSA 22

The SampleOptim tool implements optimization on age sampling based on the growth parameters, the age-length key (ALK), the mean age-at-length, and the mean length-at-age, using various statistics (see Section 2.3 and Annex 3A for detailed description of the tool).

The first step was to setup the input files with the following settings:

Table 5.6.3.1.1 - Settings used in the SampleOptimRDB for the age sampling optimization of the Red Mullet stock in GSA 22

<table>
<thead>
<tr>
<th>species</th>
<th>Mullus barbatus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AREA</td>
<td>GSA22</td>
</tr>
<tr>
<td>VARIABLE</td>
<td>age</td>
</tr>
<tr>
<td>PORT</td>
<td>FALSE</td>
</tr>
<tr>
<td>distUniPorto</td>
<td>FALSE</td>
</tr>
<tr>
<td>TIME_STRATA</td>
<td>A</td>
</tr>
<tr>
<td>SEX_RATIO</td>
<td>FALSE</td>
</tr>
<tr>
<td>MIN_LC</td>
<td>50</td>
</tr>
<tr>
<td>MAX_LC</td>
<td>260</td>
</tr>
<tr>
<td>interval_LC</td>
<td>10</td>
</tr>
<tr>
<td>MIN_age</td>
<td>0</td>
</tr>
<tr>
<td>MAX_age</td>
<td>12</td>
</tr>
<tr>
<td>MIN_OTOL.Read</td>
<td>5</td>
</tr>
<tr>
<td>MAX_OTOL.Read</td>
<td>80</td>
</tr>
<tr>
<td>interval_OTOL.Read</td>
<td>5</td>
</tr>
<tr>
<td>Linf</td>
<td>240</td>
</tr>
<tr>
<td>K</td>
<td>0.3</td>
</tr>
<tr>
<td>to</td>
<td>-1</td>
</tr>
<tr>
<td>year_start</td>
<td>2014</td>
</tr>
<tr>
<td>year_end</td>
<td>2016</td>
</tr>
<tr>
<td>n</td>
<td>100</td>
</tr>
<tr>
<td>stage_mature</td>
<td>1</td>
</tr>
</tbody>
</table>
Thus, the time stratification is annual (TIME_STRATA=A), and there is no stratification for Port (PORT= FALSE, distUniPorto= FALSE) and for sex (SEX_RATIO= FALSE). The number of age readings by length class (LC) is 5-80 (MIN_OTOL.Read- MAX_OTOL.Read), with an interval (interval_OTOL.Read) of 5.

The data used for the analysis is the CA table (SDEF) of the stock. The following Figure 5.6.3.1.1 gives an overview of the main variables used as input for the analyses.

<table>
<thead>
<tr>
<th>No</th>
<th>Variable</th>
<th>Stats / Values</th>
<th>Freqs (% of Valid)</th>
<th>Graph</th>
<th>Valid</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>month</td>
<td>Mean (sd) : 6.1 (2.4) min &lt; med &lt; max:</td>
<td>12 distinct values</td>
<td></td>
<td>2596</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 &lt; 6 &lt; 12 IQR (CV) : 2 (0.4)</td>
<td></td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td>2</td>
<td>year</td>
<td>Mean : 2015.2</td>
<td>2014: 1099 (42.3%)</td>
<td></td>
<td>2596</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Max : 2016</td>
<td>2016: 1497 (57.7%)</td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td>3</td>
<td>Length_class</td>
<td>Mean (sd) : 152 (30.4) min &lt; med &lt; max:</td>
<td>22 distinct values</td>
<td></td>
<td>2596</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50 &lt; 150 &lt; 260 IQR (CV) : 40 (0.2)</td>
<td></td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td>4</td>
<td>Weight</td>
<td>Mean (sd) : 47.8 (32) min &lt; med &lt; max:</td>
<td>957 distinct values</td>
<td></td>
<td>2596</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.9 × 39 &lt; 566.4 IQR (CV) : 35 (0.7)</td>
<td></td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td>5</td>
<td>Sex</td>
<td>1. F</td>
<td>1175 (45.3%)</td>
<td></td>
<td>2596</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. M</td>
<td>778 (30.0%)</td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. U</td>
<td>643 (24.8%)</td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td>6</td>
<td>Age</td>
<td>Mean (sd) : 2.2 (1.6) min &lt; med &lt; max:</td>
<td>12 distinct values</td>
<td></td>
<td>1338</td>
<td>1258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 &lt; 2 &lt; 12 IQR (CV) : 2 (0.7)</td>
<td></td>
<td></td>
<td>(51.54%)</td>
<td>(48.46%)</td>
</tr>
</tbody>
</table>

Figure 5.6.3.1.1 - Overview of the main variables used as input for the analyses implemented with the SampleOptimRDB tool for the Red Mullet stock in GSA 22.

The exploratory data analyses were performed using the “1_Data_exploratory_analysis_RDB.R” script of the SampleOptimRDB tool.

The output included the age and length distribution of the sampled individuals with age readings by year (Figures 5.6.3.1.2 and 5.6.3.1.3, respectively).
Subsequently, the simulations for the varying number of age readings per length class were run using the “2_Simulations_RDB_2905.R” script of the tool, and the results were produced with the “3_Simulations_results_data_analysis_RDB.R” script.
Figure 5.6.3.1.4 - Estimated parameters (A) K, (B) Linf, (C) t0 of the von Bertalanffy distribution by varying the number of age readings per LC. The cross at "-1" denotes the estimate obtained with the original dataset, with the error bars representing 1 standard error.

Figure 5.6.3.1.5 - Estimated CV of the length at age by varying the number of age readings per LC for 2014 (left), and 2016 (right).
5.6.3.2 Interpretation of results

The above results based on the von Bertalanffy growth model and the length at age suggest that a number of 30–35 age readings per length class would not compromise the estimation of the related parameters. Thus, a reduction of 30–60% could be implemented for the intermediate length classes (120–200 cm), which are also the most abundant in the fishery. For the upper and lower length classes, it seems that a significant increase in age readings may be required (see Table 5.6.2.1).

5.6.3.3 BioSim for the Red Mullet stock in GSA 22

BioSim Tool considers the ALK for the age sampling optimization, and thus, relies on the CA table.

The tool performs simulations resampling individual age measurements according to different sample size by length class (scenarios) from the original sampling for a given number of iteration and saves the age structure, the CV and the EDM (Earth Mover Distance) that can be compared with the original sample.
By principle this tool is designed to optimize the number of measurements obtained by trip. Given the stratification employed in our case, whereby the otoliths are collected annually using quota sampling by LC (as described above), the following modification was made: trips were renamed so that there would be a single trip code by year ("trip_2014" or "trip_2016"). This forces the tool to implement the optimization on the annual scale. This modification was performed in the "A_data_preparation.R" script.

The analyses were then performed using the “E_data_simulation_AGE.R” script, with the following settings, in order to allow stratification only by LC and simulate 5-80 age readings per group:

```r
species_name <- "Mullus barbatus"
short_name <- "MUT"

GSA <- "ALL"

# set sampling design of sample data
sampling_design <- list (stratified = TRUE, strata_var = "lenCls")

# expliciting the number of iterations and the sample sizes (numbers of readings by length class)
n_sims<-30
samp_sizes<-c(seq(from=5, to=80, by=5))

# Sampling of different number of individuals without replacement (sample size dependent of size classes in the sample)
sampling_options <- list (n_sims = n_sims,
  stages="one",           # no of stages
  stratification details
  stratified = TRUE,
  strata_var = "lenCls",
  stage1_samp_size=NA,
  # samp sizes
  samp_sizes = samp_sizes,
  # replacement options
  replacement=FALSE,
  sample_all_available = TRUE,
  sample_all_available_warning = FALSE,       #
  vars_to_keep = c( "Age"))

# stratification of sampleId definition
by_cat="N"
by_EUlev6="N"
by_EUlev4="N"
by_quart="N"
```
The following results were obtained.

**Figure 5.6.3.3.1** - ALK calculated as average across the iterations in all years for varying number of age readings per LC. The blue line represents the simulated and the red line the observed in the original sampling.

**Figure 5.6.3.3.2** - MWCV % calculated as average across the iterations of the overall ALK (pooled for years). The red line indicates the value of the MWCV% calculated on the original sample (baseline) and the total number of the individuals for which the age was read is reported. The dotted blue lines represent the percentiles (5th, 25th, median, 75th and 95th) estimated on the overall simulations.
5.6.3.4 Interpretation of results

The above results based on the ALK suggest that a minimum number of 30–35 age readings per length class could be considered the optimum. It is notable that the results obtained using the BioSim and the SampleOptimRDB tools are remarkably close in terms of the interpretation.

5.6.3.5 LanAge_opt for the Red Mullet stock in GSA 22

This tool allows for evaluation of the age sampling optimization effect on the age distribution of Landings for multiple years and gears, in a single area (for more info see Annex 7).

In the present case, the age sampling is stratified only by LC, thus the following settings were used:

```r
# USER INPUT -----------------------------------------------
# Select species
sppName <- "Mullus barbatus"
# select area
area = "GSA22"
years_AgStr <- c(2014, 2016) # years for age struc
years_ca <- c(2014:2016) # years for CA data (to sample ALK)
GEAR <- c("OTB_DEF", "GTR_DEF", "GNS_DEF") # select gears for age struc ...
# for all use "ALL", as they appear in the sel.gear.col
sel.gear.col <- "foCatEu6" # column for GEAR subset - should be the same as in
technical Strata in strD below
# proportion of new age.read to current no. (applied to all LCs)
props = seq(0.3, 0.9, by = 0.2)
# set stratification for Age structure
strD <- strIni(timeStrata = "quarter", techStrata = sel.gear.col)

## AGE sampling stratification
##
##
techStr_age = F # if the age data collection is stratified by the
# techStrata in strD, set this to TRUE (otherwise FALSE)
timeStr_age = F # if the age data collection is stratified by the
# timeStrata in strD, set this to TRUE (otherwise FALSE)
age_by_sex = F # is age sampling stratified for sex? (T- F)
type = "fixedK" # 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"
n.sim = 50 # no. simulations for each scenario(prop)
```

The results obtained for the present case study are shown in Figures 5.6.3.5.1–5.6.3.5.3.
Figure 5.6.3.5.1 - Landings age distribution for the Red Mullet stock in GSA 22, by gear and quarter-year. The boxplots represent the proportional reduction in the age readings, applied to all length classes. The thick black error bars represent the confidence interval of the baseline scenario.
Figure 5.6.3.5.2 - 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. (EMD: Earth Mover Distance, HEL: Hellinger distance, KS: Kolmogorov-Smirnoff distance, KUL: Kullback-Leibler distance)
Figure 5.6.3.5.3 - Distance statistics (normalized to the maximum) between the true (original) and the optimized age distribution by stratum (ie gear-quarter-year) of the landings, plotted against the fraction of reduction in the age readings by LC. (EMD: Earth Mover Distance, HEL: Hellinger distance, KS: Kolmogorov-Smirnoff distance, KUL: K)
5.6.3.6 Interpretation of results
Regarding the distance statistics (Figure 5.6.3.5.2), it seems that EMD and Kullback-Leibler distance give more clear and comparable results. Thus, it is shown that the number of age readings per LC could be reduced considerably, and a fraction of 60–70% (hence a reduction of 30–40%) would still provide similar estimates in terms of the overall age structure of the landings. Figures 5.6.3.5.1 and 5.6.3.5.3 allow for a more detailed inspection of the results by gear-year-quarter and age class. It is shown that the age classes 1–6, being the more abundant in the fishery, are mostly affected (Figure 5.6.3.5.1), and thus the sampling design could give more importance to these. In addition, it is evident that the 2016 results are more sensitive to the optimization (Figure 5.6.3.5.3). Therefore, a longer time series is required for the safer interpretation of the output.
6 Conclusions and future work

The new EU requirements in terms of data collection and advice (e.g., statistically sound sampling schemes, determination of MSY reference points for data limited stocks, bycatch of protected species recordings in at-sea sampling, collection of multispecies variables to use in Integrated Ecosystem Assessments) and temporary or permanent budget limitations in many fisheries labs are the main drivers for the optimization of the sampling protocols for biological data. From a cost/benefit perspective it is also not defensible that human resources and funds are used to collect more data than what is judged reasonably necessary for the purpose at hand. It is WKBIOPTIM’s objective to provide tools for institutes to evaluate strategies for the optimization of their biological sampling programmes (by decreasing species oversampling and/or increasing under sampling species), namely those carried out under their EU-MAP work plans and others. WKBIOPTIM has been acting as a forum to discuss, coordinate the development and compare the optimization tools for biological variables at national and stock levels. As a result, in every workshop new national or regional developments of scripts are presented along with new inputs for quality indicators, giving a valuable and constant contribute to the improvement of the optimization tools being prepared.

In addition, the extensive discussions on Quality Indicators (QIs) highlighted the importance of developing a proposal for standardizing the mathematical notation used in simulations, facilitating communication and documentation, developing a code to examine the effects of re-sampling with and without replacement on precision and accuracy of the estimates and the need to prepare an R-package with the functions used to calculate the QI to allow common indicators across analyses.

Future Work plans

During the WKBIOPTIM3 priorities for future work were identified and discussed. These concern both the consolidation of the tools developed or applied in the context of the WKBIOPTIM work, and the principles and essentials for the future application and evaluation of the ToolBox to promote the 4S (Sound Statistical Sampling Schemes). Specifically, on the latter, how the tools can interact with the RDBES will be evaluated, with further considerations on the use of the tools to work also with data from other sources (e.g. DATRAS). With regards to consolidation and validation, potential improvements have been identified, and are presented by tool in Section 2, Annex 3 and Annex 7.

Furthermore, future priority areas of development and elaboration include:

- Finalization of the code documentation for the tools already in development; Discussion of post WKBIOPTIM3 developments in existing and newly developed optimization tools; Continue the testing of the tools in a variety of case-studies;
- Continue research into quality indicators of length and age frequency data by a) testing the different indicators and quality thresholds using simulations and b) preparing an R-package with the functions used to calculate them
- Compile the scripts in a concise R-Package to enable common indicators across analyses
- Continue to give support on the use of WKBIOPTIM tools with the aim of optimization of biological sampling at national/stock/regional level.
- Discuss standardized notation for documentation of resampling techniques used in the different tools; Further investigate on the effects of resampling strategy with replacement (i.e., bootstrap) and without replacement on precision and bias of optimization results.
The aforementioned points have been considered in more detail for the Proposal on the Fourth Workshop on Optimization of Biological Sampling (WKBIOPTIM 4), presented in Annex 8.
7 References


MARE/2014/19 Med&BS project Final report, 2016 - Deliverable 2.5 - Scripts with a detailed document ensuring the reproducibility of the methods and the analyses realized with a special focus on the case studies tests - F. Gontrand, T. Rouyer, 2016.


# Annex 1: List of participants

<table>
<thead>
<tr>
<th>Participants</th>
<th>Institute</th>
<th>Country</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isabella Bitetto</td>
<td>COISPA TECNOLOGIA &amp; RICERCA, Stazione Sperimentale per lo Studio delle Risorsedel Mare, Via dei Trulli, 18/20 70126, Bari, Torre a Mare</td>
<td>Italy</td>
<td><a href="mailto:bitetto@coispa.it">bitetto@coispa.it</a></td>
</tr>
<tr>
<td>Mollie Elisabeth Brooks</td>
<td>National Institute of Aquatic Resources (DTU Aqua)</td>
<td>Denmark</td>
<td><a href="mailto:molbr@aqua.dtu.dk">molbr@aqua.dtu.dk</a></td>
</tr>
<tr>
<td>Jessica Craig</td>
<td>Marine Scotland Science, Marine Laboratory, 375 Victoria Road, Aberdeen</td>
<td>UK</td>
<td><a href="mailto:Jessica.Craig@gov.scot">Jessica.Craig@gov.scot</a></td>
</tr>
<tr>
<td>Laurent Dubroca</td>
<td>Ifremer, Port-en-Bassin Station, PO Box 32, Avenue du Général De Gaulle, 14520 Port-en-Bassin</td>
<td>France</td>
<td><a href="mailto:laurent.dubroca@ifremer.fr">laurent.dubroca@ifremer.fr</a></td>
</tr>
<tr>
<td>Ana Cláudia Fernandes (Chair)</td>
<td>Instituto Português do Mar e da Atmosfera, I.P., Rua Alfredo Magalhães Ramalho, 6 1495-006 Lisboa</td>
<td>Portugal</td>
<td><a href="mailto:acfernandes@ipma.pt">acfernandes@ipma.pt</a></td>
</tr>
<tr>
<td>Patrícia Gonçalves</td>
<td>Instituto Português do Mar e da Atmosfera, I.P., Rua Alfredo Magalhães Ramalho, 6 1495-006 Lisboa</td>
<td>Portugal</td>
<td><a href="mailto:patricia@ipma.pt">patricia@ipma.pt</a></td>
</tr>
<tr>
<td>Kirsten Håkansson</td>
<td>National Institute of Aquatic Resources (DTU Aqua)</td>
<td>Denmark</td>
<td><a href="mailto:kih@aqua.dtu.dk">kih@aqua.dtu.dk</a></td>
</tr>
<tr>
<td>Annelie Hilvarsson</td>
<td>Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Lysekil</td>
<td>Sweden</td>
<td><a href="mailto:annelie.hilvarsson@slu.se">annelie.hilvarsson@slu.se</a></td>
</tr>
<tr>
<td>Gwladys Lambert</td>
<td>Centre for Environment Fisheries and Aquaculture Science (Cefas)</td>
<td>UK</td>
<td><a href="mailto:gwladys.lambert@cefas.co.uk">gwladys.lambert@cefas.co.uk</a></td>
</tr>
<tr>
<td>Danai Mantopoulou</td>
<td>HCMR - Hellenic Centre for Marine Research</td>
<td>Greece</td>
<td><a href="mailto:danaim@hcmr.gr">danaim@hcmr.gr</a></td>
</tr>
<tr>
<td>Eirini Mantzouni (Chair)</td>
<td>Fisheries Research Institute, Nea Peramos, Kavala</td>
<td>Greece</td>
<td><a href="mailto:emantzo@inale.gr">emantzo@inale.gr</a></td>
</tr>
<tr>
<td>Katja Norén</td>
<td>Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Lysekil</td>
<td>Sweden</td>
<td><a href="mailto:katja.noren@slu.se">katja.noren@slu.se</a></td>
</tr>
<tr>
<td>Ioannis Ntokos</td>
<td>HCMR - Hellenic Centre for Marine Research</td>
<td>Greece</td>
<td><a href="mailto:gdokos@hcmr.gr">gdokos@hcmr.gr</a></td>
</tr>
<tr>
<td>Nuno Prista</td>
<td>Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Lysekil</td>
<td>Sweden</td>
<td><a href="mailto:nuno.prista@slu.se">nuno.prista@slu.se</a></td>
</tr>
<tr>
<td>Dália Reis</td>
<td>DRP/SRMCT - Azores Regional Fisheries Directorate, Azores</td>
<td>Portugal</td>
<td><a href="mailto:Dalia.CC.Reis@azores.gov.pt">Dalia.CC.Reis@azores.gov.pt</a></td>
</tr>
<tr>
<td>Kostas Touloumis</td>
<td>Fisheries Research Institute, Nea Peramos, Kavala</td>
<td>Greece</td>
<td><a href="mailto:touloumisk@inale.gr">touloumisk@inale.gr</a></td>
</tr>
<tr>
<td>Sofie Vandemaele</td>
<td>ILVO, Ankerstraat 1, 8400 Oostende</td>
<td>Belgium</td>
<td><a href="mailto:Sofie.Vandemaele@ilvo.vlaanderen.be">Sofie.Vandemaele@ilvo.vlaanderen.be</a></td>
</tr>
<tr>
<td>Julia Wischnewski</td>
<td>Institute of Sea Fisheries, Palmaille 9, 22767 Hamburg-Altona</td>
<td>Germany</td>
<td><a href="mailto:julia.wischnewski@thuenen.de">julia.wischnewski@thuenen.de</a></td>
</tr>
</tbody>
</table>
## Annex 2: Agenda

<table>
<thead>
<tr>
<th>Monday 27th May</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00 – 14:30</td>
<td>Welcome and logistics</td>
<td>Chairs</td>
</tr>
<tr>
<td>14:30 – 15:00</td>
<td>Presentation of ToR’s and adoption of agenda</td>
<td>Chairs</td>
</tr>
<tr>
<td>15:00 – 15:30</td>
<td>Presentation of developments on code used for optimization on sample level</td>
<td>Nuno Prista</td>
</tr>
<tr>
<td>15:30 – 16:00</td>
<td>SampleOptim R-tool to optimize fish sampling for biological parameters</td>
<td>Patrícia Gonçalves</td>
</tr>
<tr>
<td></td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>16:30 – 17:00</td>
<td>Quantifying robustness of a length frequency distribution shape</td>
<td>Julia Wischnewski</td>
</tr>
<tr>
<td>17:00 – 17:30</td>
<td>Data exploration and biological simulation tools in FishPi2 WP2&amp;3</td>
<td>Gwladys Lambert</td>
</tr>
<tr>
<td>17:30 – 18:00</td>
<td>Wrap up of the day</td>
<td>Plenary</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuesday 28th May</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 9:30</td>
<td>Sampling optimization and R tools: the contribution of the STREAM project</td>
<td>Isabella Bitetto</td>
</tr>
<tr>
<td>9:30 – 10:30</td>
<td>TORs a &amp; c. Discuss tool integration and documentation, define subgroups and roles.</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11:00 – 12:00</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>Lunch break</td>
<td></td>
</tr>
<tr>
<td>13:00 – 15:00</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>15:30 – 16:30</td>
<td>Discussion of issues related to tool development, QIs and documentation</td>
<td>Plenary</td>
</tr>
<tr>
<td>16:30 – 17:30</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td>17:30 – 18:00</td>
<td>Wrap up of the day</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>Evening social dinner</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wednesday 29th May</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 9:30</td>
<td>Presentation of new development on code used for optimization at sampling design level, and on QIs</td>
<td>Laurent Dubroca</td>
</tr>
<tr>
<td>9:30 – 10:00</td>
<td>Quality indicators: the contribution of the STREAM project</td>
<td>Isabella Bitetto</td>
</tr>
<tr>
<td>10:00 – 10:30</td>
<td>TOR b. Present/Summarize the combination of indicators to evaluate the quality of data under different sample sizes followed by plenary discussion on QI’s</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11:00 – 12:00</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>Lunch break</td>
<td></td>
</tr>
<tr>
<td>13:00 – 14:00</td>
<td>Discussion of issues related to tool development, QIs and documentation</td>
<td>Plenary</td>
</tr>
<tr>
<td>Time</td>
<td>Activity</td>
<td>Location</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>14:00 – 15:00</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30 – 17:30</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td>17:30 – 18:00</td>
<td>Wrap up of the day</td>
<td>Plenary</td>
</tr>
</tbody>
</table>

**Thursday 30th May**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 9:30</td>
<td>Discussion of issues related to tool development, QIs and documentation</td>
<td>Plenary</td>
</tr>
<tr>
<td>9:30 – 10:30</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00 – 12:00</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:00 – 14:00</td>
<td>Presentation and discussion of output/work done by subgroups</td>
<td>Plenary</td>
</tr>
<tr>
<td>14:00 – 15:00</td>
<td>Subgroup work</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30 – 17:30</td>
<td>Report preparation</td>
<td>Plenary</td>
</tr>
<tr>
<td>17:30 – 18:00</td>
<td>Wrap up of the day</td>
<td>Plenary</td>
</tr>
</tbody>
</table>

**Friday 31st May**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 10:30</td>
<td>Presentation and discussion of output/work done by subgroups</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00 – 11:30</td>
<td>Report preparation and discussion of future work</td>
<td>Plenary</td>
</tr>
<tr>
<td>11:30 – 12:00</td>
<td>Wrap up of the meeting</td>
<td>Chairs</td>
</tr>
</tbody>
</table>
Annex 3: Documentation of developed scripts

A - SampleOptim Documentation
Patrícia Gonçalves (IPMA), email: patricia@ipma.pt

A1. 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 randomly 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 sex ratio 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.

A2. R-packages required

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

A3. Input data

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”

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

A4. Initial values

<table>
<thead>
<tr>
<th>Names.of.variables</th>
<th>Mandatory</th>
<th>Variable.options</th>
<th>Default</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>y</td>
<td></td>
<td>CODE_FAO</td>
<td></td>
</tr>
<tr>
<td>AREA</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VARIABLE</td>
<td>y</td>
<td>age; all</td>
<td>all</td>
<td>age - only statistical analysis for age; all - statistical analysis age and maturity</td>
</tr>
<tr>
<td>PORT*</td>
<td>n</td>
<td>TRUE; FALSE</td>
<td>FALSE</td>
<td>Uses Port stratification for subsampling (TRUE); Do not consider Port stratification for subsampling (FALSE)</td>
</tr>
<tr>
<td>distUniPorto</td>
<td>n</td>
<td>TRUE; FALSE</td>
<td>FALSE</td>
<td>Uniform distribution of subsamples by Port (TRUE); Randomly distribution of subsamples by Port (FALSE)</td>
</tr>
<tr>
<td>TIME_STRATA</td>
<td>y</td>
<td>A; S; T</td>
<td>A - year; S - semester; T - quarter</td>
<td></td>
</tr>
<tr>
<td>SEX_RATIO</td>
<td>y</td>
<td>0; 1; 0&lt;numeric&lt;1; FALSE</td>
<td>0 - only males; 1 - only females; 0&lt;numeric&lt;1 sexratio proportion; FALSE - not considers sexratio</td>
<td></td>
</tr>
<tr>
<td>MIN_LC</td>
<td>y</td>
<td></td>
<td>minimum length class</td>
<td></td>
</tr>
<tr>
<td>MAX_LC</td>
<td>y</td>
<td></td>
<td>maximum length class</td>
<td></td>
</tr>
<tr>
<td>interval_LC</td>
<td>y</td>
<td>numeric</td>
<td>length class step</td>
<td></td>
</tr>
<tr>
<td>MIN_age</td>
<td>y</td>
<td></td>
<td>minimum age</td>
<td></td>
</tr>
<tr>
<td>MAX_age</td>
<td>y</td>
<td></td>
<td>maximum age</td>
<td></td>
</tr>
<tr>
<td>MIN_OTOL.Read</td>
<td>y</td>
<td></td>
<td>minimum number of individuals by length class</td>
<td></td>
</tr>
<tr>
<td>Names.of.variables</td>
<td>Mandatory</td>
<td>Variable.options</td>
<td>Default</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------</td>
<td>------------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>MAX_OTOL.Read</td>
<td>y</td>
<td></td>
<td>maximum number of individuals by length class</td>
<td></td>
</tr>
<tr>
<td>interval_OTOL.Read</td>
<td>y</td>
<td></td>
<td>interval number of individuals by length class in the simulation setup</td>
<td></td>
</tr>
<tr>
<td>Linf</td>
<td>y</td>
<td></td>
<td>von Bertallanfy growth model parameter - Linf. Used as a starting value to adjust VBGM.</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>y</td>
<td></td>
<td>von Bertallanfy growth model parameter - k. Used as a starting value to adjust VBGM.</td>
<td></td>
</tr>
<tr>
<td>t0</td>
<td>y</td>
<td></td>
<td>von Bertallanfy growth model parameter - t0. Used as a starting value to adjust VBGM.</td>
<td></td>
</tr>
<tr>
<td>year_start</td>
<td>y</td>
<td></td>
<td>first year data subset to run simulations</td>
<td></td>
</tr>
<tr>
<td>year_end</td>
<td>y</td>
<td></td>
<td>last year data subset to run simulations</td>
<td></td>
</tr>
<tr>
<td>stage_mature</td>
<td>y numeric</td>
<td></td>
<td>define the maturity stages that correspond to mature stages (to allow to determine the proportion of immatures and matures)</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>y numeric</td>
<td></td>
<td>define the number of simulations (iteration runs)</td>
<td></td>
</tr>
</tbody>
</table>

* 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).

A5. Scripts files

Script file: 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.

Script file: “1_Data_exploratory_analysis_RDB.R”
Exploratory analysis of the selected dataset to simulations are performed. The main analyses 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.
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 (sex ratio). 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 analyzed 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.

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.

A6. Outputs

An example of some of the main outputs visual and summary data tables produced by Sample-Optim are presented in the next sections.
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 $L_{\infty}$, 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 $t_0$, 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.

A7. 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).

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linf,&quot;k&quot;,&quot;t0&quot;,&quot;year&quot;,&quot;ID_sim&quot;,&quot;type&quot;</td>
<td>1,44.318935651322,0.25222904321758,-2.48287194226823,2011,1,8</td>
<td>487.44.3701387917741,0.247556421627995,-2.55063787330413,2011,2,8</td>
<td>973.44.3771125541627,0.25111032990265,-2.50611052174692,2011,3,8</td>
<td>1459.44.5852060556506,0.2452254313856,-2.5336275591729,2011,4,8</td>
<td>1945.44.409266847964,0.247295576927735,-2.54834608959638,2011,5,8</td>
<td>2431.44.5386159496939,0.245788008964458,-2.55835206891808,2011,6,8</td>
<td>2917.44.5327373073551,0.251765575455623,-2.50108016954353,2011,7,8</td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>year,&quot;L25&quot;,&quot;L50&quot;,&quot;L75&quot;,&quot;ID_sim&quot;,&quot;type&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1,2011,12.4,14.5,16.6,1.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2,2011,12.4,14.5,16.6,2.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3,2011,12.4,14.5,16.6,3.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4,2011,12.4,14.5,16.6,4.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5,2011,12.4,14.5,16.6,5.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6,2011,12.4,14.5,16.6,6.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7,2011,12.4,14.5,16.6,7.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8,2011,12.4,14.5,16.6,8.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9,2011,12.4,14.5,16.6,9.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10,2011,12.4,14.5,16.6,10.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11,2011,12.4,14.5,16.6,11.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>12,2011,12.4,14.5,16.6,12.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>13,2011,12.4,14.5,16.6,13.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>14,2011,12.4,14.5,16.6,14.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>15,2011,12.4,14.5,16.6,15.15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A8. Case studies examples

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”.
Dataset description

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

<table>
<thead>
<tr>
<th>Names.of.variables</th>
<th>Mandatory</th>
<th>Variable.values</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>y</td>
<td>MAC</td>
<td>CODE_FAO</td>
</tr>
<tr>
<td>AREA</td>
<td>y</td>
<td>27.9.a</td>
<td></td>
</tr>
<tr>
<td>VARIABLE</td>
<td>y</td>
<td>all</td>
<td>age - only statistical analysis for age; all - statistical analysis age and maturity</td>
</tr>
<tr>
<td>PORT</td>
<td>n</td>
<td>TRUE</td>
<td>Uses Port stratification for subsampling (TRUE); Do not consider Port stratification for subsampling (FALSE)</td>
</tr>
<tr>
<td>distUniPorto</td>
<td>n</td>
<td>TRUE</td>
<td>Uniform distribution of subsamples by Port (TRUE); Randomly distribution of subsamples by Port (FALSE)</td>
</tr>
<tr>
<td>TIME_STRATA</td>
<td>y</td>
<td>T</td>
<td>A - year; S - semester; T - quarter</td>
</tr>
<tr>
<td>SEX_RATIO</td>
<td>y</td>
<td>0.5</td>
<td>0 - only males; 1 - only females; 0&lt;numeric&lt;1 sex ratio proportion; FALSE - not considers sex ratio</td>
</tr>
<tr>
<td>MIN_LC</td>
<td>y</td>
<td>13</td>
<td>minimum length class</td>
</tr>
<tr>
<td>MAX_LC</td>
<td>y</td>
<td>49</td>
<td>maximum length class</td>
</tr>
<tr>
<td>interval_LC</td>
<td>y</td>
<td>1</td>
<td>length class step</td>
</tr>
<tr>
<td>MIN_age</td>
<td>y</td>
<td>0</td>
<td>minimum age</td>
</tr>
<tr>
<td>MAX_age</td>
<td>y</td>
<td>10</td>
<td>maximum age</td>
</tr>
<tr>
<td>MIN_OTOL.Read</td>
<td>y</td>
<td>1</td>
<td>minimum number of individuals by length class</td>
</tr>
<tr>
<td>MAX_OTOL.Read</td>
<td>y</td>
<td>20</td>
<td>maximum number of individuals by length class</td>
</tr>
<tr>
<td>interval_OTOL.Read</td>
<td>y</td>
<td>1</td>
<td>interval number of individuals by length class</td>
</tr>
<tr>
<td>Linf</td>
<td>y</td>
<td>50</td>
<td>von Bertallanfy growth model parameter - Linf. Used as a starting value to adjust VBGM.</td>
</tr>
<tr>
<td>K</td>
<td>y</td>
<td>0.1</td>
<td>von Bertallanfy growth model parameter - k. Used as a starting value to adjust VBGM.</td>
</tr>
<tr>
<td>t0</td>
<td>y</td>
<td>-3</td>
<td>von Bertallanfy growth model parameter - t0. Used as a starting value to adjust VBGM.</td>
</tr>
<tr>
<td>year_start</td>
<td>y</td>
<td>2011</td>
<td>first year data subset to run simulations</td>
</tr>
<tr>
<td>year_end</td>
<td>y</td>
<td>2014</td>
<td>last year data subset to run simulations</td>
</tr>
<tr>
<td>stage_mature</td>
<td>y</td>
<td>&gt;2</td>
<td>define the maturity stages that correspond to mature stages (to allow to determine the proportion of immatures and matures)</td>
</tr>
<tr>
<td>n</td>
<td>y</td>
<td>100</td>
<td>define the number of simulations (iteration runs) define the number of simulations (bootstrap runs)</td>
</tr>
</tbody>
</table>
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.

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 5.6.

A9. 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.

A10. 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.

A11. Github links

https://github.com/gonpatricia/SampleOptimRDBformat
https://github.com/ices-eg/wk_WKBIOPTIM3//SampleOptimRDB
B - SDTool 2.0

Isabella Bitetto (COISPA), email: bitetto@coispa.it

B1. Introduction

In the project MARE/2014/19 Med&BS a Sampling Design tool based on bootstrap technique (Deliverable 2.5 by F. Gontrand and T. Rouyer) was developed in R language to explore the consequences in terms of variation of precision (Coefficient of Variation (CV)) of several changes in the current sampling design (e.g. trips). This was achieved through the resampling of historical data. The underlying statistical principle is that the value of CV decreases with the increase of the number of sampling units, defining a curve. In the Sampling Design tool, the part of the curve where the tangent changes and begins to flatten (i.e. the curvature range) is considered as a suitable trade-off between precision and sampling effort and proposed as “optimal” sample size (in terms of number of trips).

In the STREAM project (Strengthening Regional cooperation in the area of fisheries biological data collection in the Mediterranean and Black Sea) the SD Tool R scripts developed in the previous grant were generalized to allow the optimization also by GSA, by area (e.g. combination of GSAs), country and gear (level 4). Moreover, new scripts to compare and evaluate the precision of different sampling stratifications were developed aimed at supporting the identification of the best trade-off between sampling effort (in terms of number of trips) and precision (CV) by including in the optimization process several inputs from expert knowledge (i.e. a constraint of at least one trip by month).

B2. Libraries required

The libraries needed to run the scripts are the following: rlist, dplyr, parallel, ggplot2, reshape2, emdist, reshape, knitr, pander, data.table.

B3. Input data

The data to be used to feed SDTool are the sampling data by primary unit (e.g. trip/haul), by species, commercial category (not mandatory) and length.

B3.1 Dataset format

The input datasets needed to run the SD Tool 2.0 must be in the COST R objects format. The CA table can also be empty, because in the analysis is not used.
CS COST object

# TR table (table of trips)

```r
head(cs_all@tr)
##   sampType landCtry vslFlgCtry year           proj trpCode vslLen vslPwr
## 1      M     BGR     BGR     2016 STREAM project 6_BGR     NA     NA
## 13     M     BGR     BGR     2016 STREAM project 3_BGR     NA     NA
## 25     M     BGR     BGR     2016 STREAM project 4_BGR     NA     NA
## 37     M     BGR     BGR     2016 STREAM project 7_BGR     NA     NA
## 49     M     BGR     BGR     2016 STREAM project 2_BGR     NA     NA
## 61     M     BGR     BGR     2016 STREAM project 8_BGR     NA     NA
##   vslSize vslType harbour foNum daysAtSea vslId sampCtry sampMeth
## 1     NA     <NA>    <NA>   -1     -1     1     NA     <NA> Observer
## 13    NA     <NA>    <NA>   -1     -1     1     NA     <NA> Observer
## 25    NA     <NA>    <NA>   -1     -1     1     NA     <NA> Observer
## 37    NA     <NA>    <NA>   -1     -1     1     NA     <NA> Observer
## 49    NA     <NA>    <NA>   -1     -1     1     NA     <NA> Observer
## 61    NA     <NA>    <NA>   -1     -1     1     NA     <NA> Observer
```

# HH table (table of fishing operations/hauls)

```r
head(cs_all@hh)
##   sampType landCtry vslFlgCtry year           proj trpCode staNum foVal
## 1      M     BGR     BGR     2016 STREAM project 6_BGR     1     V
## 13     M     BGR     BGR     2016 STREAM project 3_BGR     1     V
## 25     M     BGR     BGR     2016 STREAM project 4_BGR     1     V
## 37     M     BGR     BGR     2016 STREAM project 7_BGR     1     V
## 49     M     BGR     BGR     2016 STREAM project 2_BGR     1     V
## 61     M     BGR     BGR     2016 STREAM project 8_BGR     1     V
##   aggLev catReg sppReg       date time foDur latIni lonIni latFin lonFin
## 1     T    All    Par 15/06/2016 <NA>   -1     -1     -1     -1     -1
## 13    T    All    Par 12/07/2016 <NA>   -1     -1     -1     -1     -1
## 25    T    All    Par 12/07/2016 <NA>   -1     -1     -1     -1     -1
## 37    T    All    Par 16/08/2016 <NA>   -1     -1     -1     -1     -1
## 49    T    All    Par 08/09/2016 <NA>   -1     -1     -1     -1     -1
## 61    T    All    Par 18/10/2016 <NA>   -1     -1     -1     -1     -1
##   area rect subRect foDep waterDep foCatNat foCatEu5 foCatEu6 meshSize
## 1 GSA29 <NA>    <NA>   -1     -1     -1     OTM_MPD OTM_MPD NA
## 13 GSA29 <NA>    <NA>   -1     -1     -1     OTM_MPD OTM_MPD NA
## 25 GSA29 <NA>    <NA>   -1     -1     -1     OTM_MPD OTM_MPD NA
## 37 GSA29 <NA>    <NA>   -1     -1     -1     OTM_MPD OTM_MPD NA
## 49 GSA29 <NA>    <NA>   -1     -1     -1     OTM_MPD OTM_MPD NA
## 61 GSA29 <NA>    <NA>   -1     -1     -1     OTM_MPD OTM_MPD NA
##   selDev meshSizeSelDev
## 1     NA
## 13    NA
## 25    NA
```
# SL table (table of species lists/samples)

```r
head(cs_all@sl)
##         sampType landCtry vslFlgCtry year           proj trpCode staNum
## 1          M      BGR        BGR 2016 STREAM project   6_BGR      1
## 13         M      BGR        BGR 2016 STREAM project   3_BGR      1
## 25         M      BGR        BGR 2016 STREAM project   4_BGR      1
## 37         M      BGR        BGR 2016 STREAM project   7_BGR      1
## 49         M      BGR        BGR 2016 STREAM project   2_BGR      1
## 61         M      BGR        BGR 2016 STREAM project   8_BGR      1
##                          spp catchCat landCat commCatScl commCat subSampCat sex   wt
## 1 Mullus barbatus      LAN     HUC -1 -1 -1 -1 1431
## 13 Mullus barbatus     LAN     HUC -1 -1 -1 -1 1367
## 25 Mullus barbatus     LAN     HUC -1 -1 -1 -1 1455
## 37 Mullus barbatus     LAN     HUC -1 -1 -1 -1 1363
## 49 Mullus barbatus     LAN     HUC -1 -1 -1 -1 1685
## 61 Mullus barbatus     LAN     HUC -1 -1 -1 -1 1638
##                          subSampWt lenCode
## 1       1431      mm
## 13      1367      mm
## 25      1455      mm
## 37      1363      mm
## 49      1685      mm
## 61      1638      mm
```

# HL table (table of the length distributions by sample)

```r
head(cs_all@hl)
##         sampType landCtry vslFlgCtry year           proj trpCode staNum
## 1          M      BGR        BGR 2016 STREAM project   1_BGR      1
## 2          M      BGR        BGR 2016 STREAM project   5_BGR      1
## 3          M      BGR        BGR 2016 STREAM project   3_BGR      1
## 4          M      BGR        BGR 2016 STREAM project   7_BGR      1
## 5          M      BGR        BGR 2016 STREAM project   3_BGR      1
## 6          M      BGR        BGR 2016 STREAM project   4_BGR      1
##                          spp catchCat landCat commCatScl commCat subSampCat sex
## 1 Mullus barbatus      LAN     HUC -1 -1 -1 -1
## 2 Mullus barbatus      LAN     HUC -1 -1 -1 -1
## 3 Mullus barbatus      LAN     HUC -1 -1 -1 -1
## 4 Mullus barbatus      LAN     HUC -1 -1 -1 -1
## 5 Mullus barbatus      LAN     HUC -1 -1 -1 -1
## 6 Mullus barbatus      LAN     HUC -1 -1 -1 -1
##                          lenCls lenNum
## 1     90   1
```
## CA table (table of the individuals measures)

```r
head(cs_all@ca)
## sampType landCtry vslFlgCtry year proj trpCode staNum quarter month spp
## 1 <NA> <NA> <NA> NA <NA> <NA> NA NA NA <NA>
## 2 <NA> <NA> <NA> <NA> <NA> <NA> NA NA NA <NA>
## 3 <NA> <NA> <NA> <NA> <NA> <NA> <NA> NA NA <NA>
## 4 <NA> <NA> <NA> <NA> <NA> <NA> <NA> NA NA <NA>
## 5 <NA> <NA> <NA> <NA> <NA> <NA> <NA> NA NA <NA>
## 6 <NA> <NA> <NA> <NA> <NA> <NA> <NA> NA NA <NA>

# sex catchCat landCat commCatScl commCat stock area rect subRect lenCls
## 1 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 2 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 3 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 4 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 5 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 6 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>

# age fishId lenCode ageMeth plusGrp otoWt otoSide indWt matMeth matScale
## 1 NA NA NA NA NA NA NA NA NA NA
## 2 NA NA NA NA NA NA NA NA NA NA
## 3 NA NA NA NA NA NA NA NA NA NA
## 4 NA NA NA NA NA NA NA NA NA NA
## 5 NA NA NA NA NA NA NA NA NA NA
## 6 NA NA NA NA NA NA NA NA NA NA

# matStage
## 1 <NA>
```

## CL COST object

## CL table (table of the commercial landings)

```r
head(cl_all@cl)
## landCtry vslFlgCtry year quarter month area rect subRect
## 1 ITA ITA 2014 1 NA <NA> <NA>
## 2 ITA ITA 2014 1 NA <NA> <NA>
## 3 ITA ITA 2014 2 NA <NA> <NA>
## 4 ITA ITA 2014 2 NA <NA> <NA>
## 5 ITA ITA 2014 3 NA <NA> <NA>
## 6 ITA ITA 2014 3 NA <NA> <NA>

# taxon landCat commCatScl commCat foCatNat foCatEu5
## 1 Merluccius merluccius <NA> <NA> <NA> <NA> <NA>
## 2 Merluccius merluccius <NA> <NA> <NA> <NA> <NA>
## 3 Merluccius merluccius <NA> <NA> <NA> <NA> <NA>
## 4 Merluccius merluccius <NA> <NA> <NA> <NA> <NA>
## 5 Merluccius merluccius <NA> <NA> <NA> <NA> <NA>
## 6 Merluccius merluccius <NA> <NA> <NA> <NA> <NA>

# foCatEu6 harbour vslLenCat unallocCatchWt misRepCatchWt
## 1 OTB_DEF_>=40_0_0 <NA> <NA> NA NA
## 2 OTB_DWS_>=40_0_0 <NA> <NA> NA NA
## 3 OTB_DEF_>=40_0_0 <NA> <NA> NA NA
## 4 OTB_DWS_>=40_0_0 <NA> <NA> NA NA
## 5 OTB_DEF_>=40_0_0 <NA> <NA> NA NA
## 6 OTB_DWS_>=40_0_0 <NA> <NA> NA NA

# landWt landMult landValue
## 1 307099.01 NA 1738460
## 2 1962.40 NA 12406
```
### Description of the variables of the dataset

For the description of the variables stored according the COST format, refer to Jansen (2009).

### Initial values/starting values for simulations

A dummy dataset is available on GitHub at: https://github.com/aleligas/STREAM_MARE-2016-22/tree/Task-3.3---Sampling-Strategy-Optimization.

### Scripts files

The SD tool consists of the following 4 main scripts written in R language, saved in `.Rmd` format, to be run in R studio (tested in R studio version 1.1.456 with R version 3.3.1):

- **04_investigateData.Rmd** for investigation on historical sampling data;
- **05_1_runOptimizationBYspecies.Rmd** performing the analyses for sampling optimization in terms of number of trips by species based on the analyses of the CV. The script return multiple solutions to be inspected by the user (not only one range of 2 solutions);
- **05_2_runOptimizationBYspecies.Rmd** allowing the production of some plots comparing the current sampling with the optimal size range selected by the user among the provided solutions in the previous step;
- **06_runScenario.Rmd** performing simulations based on a fixed number of trips (scenarios of type from 1 to 3) and, eventually, decrease of the number of length measurements;
- **07_runEvaluation.Rmd** performing the comparison of the different simulations assessing the increase/decrease in CV compared to the current situation and implementing the calculation of EMD (Earth Mover’s Distance);
- **utility_appendOptimizations.Rmd** appending the results related to the optimal sample size obtained for each target species.

A folder named `SDTfuns` contains auxiliary functions implementing single steps of the algorithms re-called by the main scripts listed above.
Figure B.1 – Configuration of the SD Tool 2.0 main folder with main scripts and auxiliary functions.

**B4.1 Script 04_investigateData.Rmd**

- Compile the settings in the script *04_investigateData.Rmd* indicating the following elements:
  - `myWD`: the path of the working directory, i.e. the path where SD Tool 2.0 folder is located;
  - `results_path`: the path of the directory of the case study where the input files are located and all the outputs will be saved in the related folders (created by the tool);
  - `cl_all`: the path where the COST CL object is located must be set in the code;
  - `cs_all`: the path where the COST CS object is located must be set in the code;
  - `table_strat_res`: the path of the stratification results table (see Table B.2)
  - Table B.2 target_SPECIES: the list of the target species with a short code (e.g. "DPS") and the scientific name reported in the CS and CL datasets (e.g. "Parapenaeus longirostris")

```r
# set the working directory
# set the working directory
myWD <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19\SCRIPTS\SD Tool 2.0"
setwd(myWD)
results_path <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19"

cl_all <- readRDS(paste(results_path, "/input files/costCL.rds", sep=""))
cs_all <- readRDS(paste(results_path, "/input files/costCS.rds", sep=""))

table_strat_res <- read.csv(paste(results_path, "/input files/05_stratification_results.csv", sep=""), sep=";")

target_SPECIES <- list("DPS" = "Parapenaeus longirostris",
                      "ARA" = "Aristeus antennatus")
```
Use the `stratification_results.csv` file (see Table B.2) to define the level of aggregation of the current sampling scheme for space, time and technical level of the fishing technique according DCF. It could be considered the criterion used to define the sampling domains, e.g. `GSA-year-lev6`, `GSA-year-lev4`, `Country-year-lev4`. From a spatial point of view, the analyses can be performed by GSA, by Country or all the available areas/countries together (`ALL`); from a temporal point of view the analyses can be performed by year, by quarter or by semester; as regards the technical stratification, the analyses can be performed by gear (`level4`), by métier (`level6`) or all the available métiers together (`NONE`).

Table B.2 - Structure of the stratification results table used to define the level of aggregation of the sampling in terms of spatial, temporal and technical unit. In the table only one of the items listed in the rows can be inserted to define the level of aggregation (e.g. `GSA-year-lev6`).

<table>
<thead>
<tr>
<th>space</th>
<th>time</th>
<th>technical</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSA</td>
<td>Y</td>
<td>lev4</td>
</tr>
<tr>
<td>Country</td>
<td>Q</td>
<td>lev6</td>
</tr>
<tr>
<td>ALL</td>
<td>S</td>
<td>NONE</td>
</tr>
</tbody>
</table>

- Open script #4 in R studio and run it with knit button to perform investigation on historical sampling data.
- The outputs tables in `.csv` format and the plots are saved by the script in the folder `04_dataInvestigation` (see from Figure B.2 – example Table B.3 to Table B.8 and from Figure B.3 to Figure B.5).
Figure B.2 – Example of delta values by quarter.

Table B.3 - Structure of the table containing the values of the variability between samples by metier (separate files by species and GSA).

<table>
<thead>
<tr>
<th>Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Samp</td>
<td># trip</td>
</tr>
<tr>
<td>Delta</td>
<td>$\Delta$ calculated according Vigneau &amp; Mahévas (2007)</td>
</tr>
<tr>
<td>tc</td>
<td>Metier short code (e.g. OTB_MDD)</td>
</tr>
<tr>
<td>trpCode</td>
<td>Trip ID code</td>
</tr>
<tr>
<td>staNum</td>
<td>Station number (haul)</td>
</tr>
<tr>
<td>spp</td>
<td>Species</td>
</tr>
</tbody>
</table>
Table B.4 - Structure of the table containing the values of the variability between samples by quarter (separate files by species and GSA).

<table>
<thead>
<tr>
<th>Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>samp # trip</td>
<td></td>
</tr>
<tr>
<td>delta Δ calculated according Vigneau &amp; Mahévas (2007)</td>
<td></td>
</tr>
<tr>
<td>tp Quarter, as reported in the dataset (e.g. 1, 2, ...)</td>
<td></td>
</tr>
<tr>
<td>trpCode Trip ID code</td>
<td></td>
</tr>
<tr>
<td>staNum Station number (haul)</td>
<td></td>
</tr>
<tr>
<td>spp Species</td>
<td></td>
</tr>
</tbody>
</table>

Table B.5 - Structure of the table containing the values of the variability between samples by year (separate files by species and GSA).

<table>
<thead>
<tr>
<th>Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>samp # trip</td>
<td></td>
</tr>
<tr>
<td>delta Δ calculated according Vigneau &amp; Mahévas (2007)</td>
<td></td>
</tr>
<tr>
<td>tp Year, as reported in the dataset (e.g. 2015)</td>
<td></td>
</tr>
<tr>
<td>trpCode Trip ID code</td>
<td></td>
</tr>
<tr>
<td>staNum Station number (haul)</td>
<td></td>
</tr>
<tr>
<td>spp Species</td>
<td></td>
</tr>
</tbody>
</table>

Table B.6 - Structure of the table containing the values of the landings by year-quarter-foCatEu6 (unique file for all the species and GSAs).

<table>
<thead>
<tr>
<th>Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Area GSA, as reported in the dataset (e.g. GSA18)</td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td></td>
</tr>
<tr>
<td>Quarter Quarter, as reported in the dataset (e.g. 1, 2, ...)</td>
<td></td>
</tr>
<tr>
<td>Year Year</td>
<td></td>
</tr>
<tr>
<td>foCatEu6 Metier short code (e.g. OTB_MDD)</td>
<td></td>
</tr>
<tr>
<td>Value Landing in kg</td>
<td></td>
</tr>
</tbody>
</table>
Table B.7. Structure of the table containing the number of trips-samples-measures (unique file for all the species and GSAs).

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
</tr>
<tr>
<td>GSA, as reported in the dataset (e.g. GSA18)</td>
</tr>
<tr>
<td>Year</td>
</tr>
<tr>
<td>Year</td>
</tr>
<tr>
<td>Metier</td>
</tr>
<tr>
<td>Metier short code (e.g. OTB_MDD)</td>
</tr>
<tr>
<td>Species</td>
</tr>
<tr>
<td>Species</td>
</tr>
<tr>
<td>TQ1</td>
</tr>
<tr>
<td>Number of trips in the 1st quarter (identified by trpCode)</td>
</tr>
<tr>
<td>TQ2</td>
</tr>
<tr>
<td>Number of trips in the 2nd quarter (identified by trpCode)</td>
</tr>
<tr>
<td>TQ3</td>
</tr>
<tr>
<td>Number of trips in the 3rd quarter (identified by trpCode)</td>
</tr>
<tr>
<td>TQ4</td>
</tr>
<tr>
<td>Number of trips in the 4th quarter (identified by trpCode)</td>
</tr>
<tr>
<td>SQ1</td>
</tr>
<tr>
<td>Number of samples in the 1st quarter (identified by trpCode, staNum and commCat)</td>
</tr>
<tr>
<td>SQ2</td>
</tr>
<tr>
<td>Number of samples in the 2nd quarter (identified by trpCode, staNum and commCat)</td>
</tr>
<tr>
<td>SQ3</td>
</tr>
<tr>
<td>Number of samples in the 3rd quarter (identified by trpCode, staNum and commCat)</td>
</tr>
<tr>
<td>SQ4</td>
</tr>
<tr>
<td>Number of samples in the 4th quarter (identified by trpCode, staNum and commCat)</td>
</tr>
<tr>
<td>NQ1</td>
</tr>
<tr>
<td>Number of measured individuals in the 1st quarter</td>
</tr>
<tr>
<td>NQ2</td>
</tr>
<tr>
<td>Number of measured individuals in the 2nd quarter</td>
</tr>
<tr>
<td>NQ3</td>
</tr>
<tr>
<td>Number of measured individuals in the 3rd quarter</td>
</tr>
<tr>
<td>NQ4</td>
</tr>
<tr>
<td>Number of measured individuals in the 4th quarter</td>
</tr>
<tr>
<td>annualT</td>
</tr>
<tr>
<td>Total number of trips in the year (identified by trpCode)</td>
</tr>
<tr>
<td>annualS</td>
</tr>
<tr>
<td>Total number of samples in the year (identified by trpCode, staNum and commCat)</td>
</tr>
<tr>
<td>annualN</td>
</tr>
<tr>
<td>Total number of measured individuals</td>
</tr>
</tbody>
</table>
Table B.8 - Structure of the Past situation table (unique file for all the species and GSAs).

<table>
<thead>
<tr>
<th>Description</th>
<th>Var1</th>
<th>Var2</th>
<th>Var3</th>
<th>x.x</th>
<th>Sampling</th>
<th>x.y</th>
<th>Landings</th>
<th>Year</th>
<th>sampSize</th>
<th>nMeasTotal</th>
<th>MeanNmeasByStr</th>
<th>Cv</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Spatial stratification (e.g. GSA)</td>
<td>Time stratification (e.g. 1, 2, 3, 4 for quarters; 1 for year)</td>
<td>Metier/gear stratification</td>
<td>Total weight (kg) of samples in the stratum</td>
<td>Proportion of the weight of the samples among the strata in the year (e.g. among quarters)</td>
<td>Commercial landings (kg) in the stratum</td>
<td>Proportion of the commercial landings among the strata in the year (e.g. among GSA 18 and GSA 19)</td>
<td>Year</td>
<td>Total number of trips in the stratum</td>
<td>Total number of measured individuals in the stratum</td>
<td>Mean number of measured individuals by trip in the stratum</td>
<td>Coefficient of Variation calculated according Vigneau &amp; Mahévas (2007)</td>
<td>Species</td>
</tr>
</tbody>
</table>
Figure B.3 - Sample weight by trip, fishing activity category European lvl 6 and quarter (one plot per year).

Figure B.4 - Relative importance of sampling compared to the relative importance of landing by fishing activity category European lvl 6 and quarter (one plot per year).
The tables and figures with delta values, reported in Table B.5 and Figure B.5, can be useful to identify outliers (i.e. delta values out of the ranges for the considered métier, quarter or year). Indeed, the variability plot with delta values can be useful to identify similar/dissimilar variability among different métiers or quarters, thus supporting the decision on the sampling effort allocation between métiers and/or quarters. Under certain circumstances (e.g. fisheries share the same targets in terms of components of a given species population), it could also help to choose the level of aggregation in the CV analyses (e.g. if the variability is similar among the métiers, it might be possible to consider the gear - lev4 - as technical aggregation in the sampling scheme, see the TableB.2).

B4.2 Script Optimize sampling size (number of trips) for the current sampling scheme (scripts #5_1 and #5_2)

- Open the script in R studio and compile the settings in the 05_1_runOptimization BYSpecies.Rmd script indicating the following elements:
  - REFERENCE_SPECIES: the reference species for the optimization;
  - RS_shortcode: a short code for the reference species to be used in the output;
  - cl_all: the path where the COST CL object is located must be set in the code;
  - cs_all: the path where the COST CS object is located must be set in the code;
- `pastSituations`: the path of the `Past situation.csv` table (see Table B.8) saved by the script #4;
- `table_strat_res`: the path of the `05_1_stratification_results.csv` table (see Table B.2);
- `nIter`: the number of iterations to perform with the bootstrap;
- `table_min_accepted_sample_size`: the path of the `05_2_min_accepted_sample_size.csv` table containing the minimum number of trips accepted for each defined stratum, e.g. 3 trips in case of stratification by quarter would ensure at least 1 trip per month (see Table B.9);
- `threshold_for_RecyclingRate`: a threshold for the recycling rate to be considered for the assessment of the reliability of the results.

```r
REFERENCE_SPECIES <- "Parapenaeus longirostris"
RS_shortcode <- "DPS"

cl_all <- readRDS(paste(CASE_STUDY_PATH, "/input files/costCL.rds", sep=""))
cs_all <- readRDS(paste(CASE_STUDY_PATH, "/input files/costCS.rds", sep=""))

pastSituations <- read.csv(paste(CASE_STUDY_PATH, "/04_dataInvestigation/Past situation.csv", sep=""), sep=";"))
table_strat_res <- read.csv(paste(CASE_STUDY_PATH, "/input files/05_stratification_results.csv", sep=""), sep=";"))
nIter <- 100
table_min_accepted_sample_size<- read.csv(paste(CASE_STUDY_PATH, "/input files/05_2_min_accepted_sample_size.csv", sep=""), sep=";"))
threshold_for_RecyclingRate <- 0.5
```

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Space</strong></td>
</tr>
<tr>
<td><strong>Time</strong></td>
</tr>
<tr>
<td><strong>Technical</strong></td>
</tr>
<tr>
<td><strong>min_accepted_ss</strong></td>
</tr>
</tbody>
</table>
• Use script 05_1_runOptimizationBYspecies.Rmd to perform analyses for sampling optimization and obtain all the solutions given by the optimization process. Then, a pair of solutions have to be selected to define the “optimal” sampling size, on the basis on the analyses of the CV (e.g. stability of CVs from bootstrap and dispersion of them) of a given species. This script includes options for a flexible definition of the sampling stratification considering: technical strata (by métier/lev6 or by gear/lev4), temporal aggregation (by year, quarter or semester), spatial level (by GSA or Country/MS level). ATTENTION: when the script is run, a window appears to allow the modification of a line of code in a COST function to fix a bug according the following indication in the code.

```r
# ATTENTION!!! # BEFORE STARTING RUN THE CODE FIX THE BUG IN THE COST FUNCTION RUNNING THE FOLLOWING CODE:

fixInNamespace(spdAgreg, pos="package:COSTdbe")

# WHEN THE WINDOWS APPEARS, replace the line 8 with the following code:
# val <- paste(unlist(BY), collapse = "-:-")
```

• Use the same 05_1_stratification_results.csv used in the data investigation phase (see Table B.2).
• Open script #5_1 with R studio and run it using knit button for each species of interest.
• The outputs tables and the plots saved in .csv format in the folder 05_optimization (for each species a different folder will be created) are reported below (see from Table B.10 to Table B.13 and from Figure B.6 to Figure B.8). The Optimized sample size.csv table reports all the solutions for each species. The reliability of the result is lower if the RR (Recycling Rate) is high (see Table B.11).

Table B.10 – Structure of the table containing the values of the Coefficient of Variation by iteration in output from the bootstrap.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cv</td>
</tr>
<tr>
<td>nMeas</td>
</tr>
<tr>
<td>sampSize</td>
</tr>
<tr>
<td>recyclingRate</td>
</tr>
<tr>
<td>Var1</td>
</tr>
<tr>
<td>Var2</td>
</tr>
<tr>
<td>Var3</td>
</tr>
</tbody>
</table>

- Cv: Coefficient of Variation calculated according Vigneau & Mahévas (2007)
- nMeas: Total number of measured individuals in the stratum
- sampSize: Total number of trips in the stratum
- recyclingRate: Percentage of re-use of the same trip when running the bootstrap
- Var1: Spatial stratification (e.g. GSA)
- Var2: Time stratification (e.g. 1, 2, 3, 4 for quarters; 1 for year)
- Var3: Metier/gear stratification
Table B.11 – Structure of the Optimized sample size table containing the first 4 solutions inferred via the method, i.e. the “optimal” number of trips greater than the minimum solution indicated by the user as acceptable for each stratum.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var1</td>
</tr>
<tr>
<td>Var2</td>
</tr>
<tr>
<td>Var3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trips corresponding to one of the local maxima found in the density function.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxCV</td>
</tr>
<tr>
<td>minCV</td>
</tr>
<tr>
<td>maxRR</td>
</tr>
<tr>
<td>minRR</td>
</tr>
<tr>
<td>meanRR</td>
</tr>
<tr>
<td>noIterations</td>
</tr>
</tbody>
</table>

Table B.12 – Structure of the table containing the Length distributions raised to the total landing from bootstrap by iteration.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>iteration</td>
</tr>
<tr>
<td>Var1</td>
</tr>
<tr>
<td>Var2</td>
</tr>
<tr>
<td>Var3</td>
</tr>
<tr>
<td>length</td>
</tr>
<tr>
<td>value</td>
</tr>
</tbody>
</table>
Figure B.6 – Density function of the CV used by the method to infer the solutions (optimal number of trips). The red lines indicate all the solutions inferred as “optimal” sampling size and for each line the number of trips and the min-max value of the CV are reported.

Figure B.7 - CV versus the number of trips. Trip codes by metier (level 4) and year. Vertical blue lines indicate the first 4 solutions in terms of number of trips (greater than the minimum accepted solution indicated by the user as acceptable) inferred via the optimization process.
Figure B.8 - Recycling rate of the samples against the number of trips by metier (level 4) and year. Vertical blue lines are referred to the first 4 solutions in terms of number of trips (greater than the minimum accepted solution indicated by the user as acceptable) inferred via the method. Red line corresponded to an arbitrary critical threshold of 50%.

- Open in R studio the script `05_2_runOptimizationBYspecies.Rmd` and run it after the definition of the “optimal” sampling size (minimum and maximum number of trips chosen from all the solutions found in the previous step) in the file sampSizeRange (Table B.13). Some additional plots are produced using the indication of the selected optimal sample size range (Figure B.9 and Figure B.10).
- Compile the settings in the `05_2_runOptimizationBYspecies.Rmd` script indicating the following elements:
  - `REFERENCE_SPECIES`: the reference species for the optimization;
  - `RS_shortcode`: a short code for the reference species to be used in the output;
  - `cl_all`: the path where the COST CL object is located must be set in the code;
  - `cs_all`: the path where the COST CS object is located must be set in the code;
  - `pastSituations`: the path of the `Past situation.csv` table (see Table B.8) saved by the script #4;
  - `table_strat_res`: the path of the `stratification_results.csv` table (see Table B.2);
  - `nIter`: the number of iterations to perform with the bootstrap;
  - `sampSizeRange`: path of the table where the user has defined “optimal” sampling size range (min and max number of trips) among the solutions inferred via the method (see Table B.13).

```r
REFERENCE_SPECIES <- "Parapenaeus longirostris"
RS_shortcode <- "DPS"

cl_all <- readRDS(paste(CASE_STUDY_PATH, "/input files/costCL.rds", sep=""))
cs_all <- readRDS(paste(CASE_STUDY_PATH, "/input files/costCS.rds", sep=""))
pastSituations <- read.csv(paste(CASE_STUDY_PATH, "/04_dataInvestigation/Past situation.csv", sep=""), sep=";")
table_strat_res <- read.csv(paste(CASE_STUDY_PATH, "/input files/05_stratification_results.csv", sep=""), sep=";")
```
```r
sampSizeRange <- read.csv(paste(CASE_STUDY_PATH, "/input files/DPS - Optimized sample size SELECTED.csv", sep=""), sep=";")
nIter <<- 100
```

Table B.13 – Structure of the table containing the chosen solutions (min and max number of trips) defining the “optimal” sampling size range.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var1</td>
</tr>
<tr>
<td>Spatial stratification (e.g. GSA)</td>
</tr>
<tr>
<td>Var2</td>
</tr>
<tr>
<td>Time stratification (e.g. 1, 2, 3, 4 for quarters; 1 for year)</td>
</tr>
<tr>
<td>Var3</td>
</tr>
<tr>
<td>Metier/gear stratification</td>
</tr>
<tr>
<td>Solutions</td>
</tr>
<tr>
<td>Number of trips corresponding to one of the local maxima found in the density function</td>
</tr>
<tr>
<td>maxCV</td>
</tr>
<tr>
<td>Maximum CV found in the range between the indicated solution and the previous one (e.g. between 24 and 30 trips) calculated on the niterations number of iterations</td>
</tr>
<tr>
<td>minCV</td>
</tr>
<tr>
<td>Minimum CV found in the range between the indicated solution and the previous one (e.g. between 24 and 30 trips) calculated on the niterations number of iterations</td>
</tr>
<tr>
<td>maxRR</td>
</tr>
<tr>
<td>Maximum recycling rates found in the range between the indicated solution and the previous one (e.g. between 24 and 30 trips) calculated on the niterations number of iterations</td>
</tr>
<tr>
<td>minRR</td>
</tr>
<tr>
<td>Minimum recycling rates found in the range between the indicated solution and the previous one (e.g. between 24 and 30 trips) calculated on the niterations number of iterations</td>
</tr>
<tr>
<td>meanRR</td>
</tr>
<tr>
<td>Average recycling rates found in the range between the indicated solution and the previous one (e.g. between 24 and 30 trips) calculated on the niterations number of iterations</td>
</tr>
<tr>
<td>niterations</td>
</tr>
<tr>
<td>Number of iterations in the range between the indicated solution and the previous one (e.g. between 24 and 30 trips)</td>
</tr>
<tr>
<td>min_max</td>
</tr>
<tr>
<td>Indicate if the solution is the minimum (min) or the maximum (max) of the “optimal” sampling size range defined in the considered stratum</td>
</tr>
</tbody>
</table>
Table B.14 – Structure of the table containing the derived no. of individuals from past sampling.

<table>
<thead>
<tr>
<th>Description</th>
<th>Var1</th>
<th>Var2</th>
<th>Var3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Spatial stratification (e.g. GSA)</td>
<td>Time stratification (e.g. 1, 2, 3, 4 for quarters; 1 for year)</td>
<td>Metier/gear stratification</td>
</tr>
<tr>
<td>Derived no. of individuals (mean)</td>
<td>Average of the number of individuals obtained from the bootstrap (n iteration) in the “optimal” sample size range estimated by the “optimization” process (e.g. between 10 and 25 trips)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Derived no. of individuals (st. dev.)</td>
<td>Standard deviation of the number of individuals obtained from the bootstrap (n iteration) in the “optimal” sample size range estimated by the “optimization” process (e.g. between 10 and 25 trips)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure B.9 - Optimal sampling size in terms of number of trips by metier (level 4) and year selected by the user among the solution inferred via the method compared to the historical sampling data. Green zone represents the optimal sample size, the yellow +/- 33% of the range to the limits and the red is +/- 66% of the range to the limits.
Figure B.10 – Plot with the derived number of individuals measured in the past sampling (x-axis) versus the number of individuals to be measured estimated by the tool taking into account the “optimal” sampling size range defined by the user through the two chosen solutions.

B4.3 Script Run Scenario
For any type of scenario, compile the settings in the script #6 indicating the following elements:

- **myWD**: the path of the working directory, i.e. the path where SD Tool 2.01 folder is located;
- **results_path**: the path of the directory of the case study where the input files are located and all the outputs will be saved in the related folders (created by the tool);
- **cl_all**: the path where the COST CL object is located must be set in the code;
- **cs_all**: the path where the COST CS object is located must be set in the code;
- **path_metier_shortcode**: the path where the table of correspondences between Lev6 and Lev4 is located;
- **path_table_strat_results**: the path of the 05_1 stratification_results.csv table (see Table B.2);
- **path_table_sampling_definition**: the path of the 06_1 sampling_design_definition.csv table reporting the level of aggregation (spatial, temporal and technical) to be used for the allocation of the trips (see Table B.15). This table can report:
  - the same level of aggregation used in the optimization process and defined in the 05_1 stratification_results.csv, e.g. if you optimized the trips using the stratification gear/lev4 and you do not want to define the percentage of trips to be allocated among the métiers, make the bootstrap select the trips randomly among the trips of all the métiers;
  - a level of aggregation more detailed than the one defined in the 05_1 stratification_results.csv table, e.g. if you want to split a number of trips by gear/lev4 in the different métier. In the abovementioned example the 06_1 sampling_design_definition.csv table should, for example report, the number of trips by métier/lev6.
The first `path_table_sampling_definition` can be used if the optimization process was already carried out at a lower level of aggregation (by year or by gear), but the simulation of more detailed stratifications should be explored; in this case, the second `path_table_sampling_definition` can be used to explore alternative allocations among the métier in combination with the file `06_2_sampling_design_trips_sharing.csv`:

- `path_table_trips_sharing`: the path of the `06_2_sampling_design_trips_sharing.csv` table containing the numbers of trips to be used to allocate the trips in each stratum (see Table B.16). Columns A, C and E should be respectively identical to the columns A, B, C in the output csv “Optimized sample size” obtained with optimization process (script #5). Use columns B, D, F, G to reallocate the trips.

- `path_table_subsamples`: the path of the `06_3_subsample_definition.csv` table containing indications on eventual subsampling to be simulated in each stratum for each species and, optionally, by commercial category: a number lower or equal to 1 that corresponds to the fraction to be used for the subsampling (e.g. 0.5 for \( \frac{1}{2} \)) and a number of individuals to be used as threshold to apply the subsampling should be indicated (see Table B.17). In particular, the following information have to be set:
  - `factor`: subsample factor for subsampling in each trip/haul (e.g. \( \frac{1}{2} \) indicates that the user wants to simulate that the sample of a given trip/haul, eventually by commercial category, is divided by 2);
  - `threshold`: the minimum number of individuals by trip accepted for a subsample. When simulating scenario with subsampling, if the number of subsampled individuals is lower than the threshold (with a buffer of 5%), the tool will not use the subsample but the sample itself. For example, if the original sample has 300 measured individuals and the subsample factor is \( \frac{1}{2} \) and the threshold is:
    - 100, the subsample is used in the simulation having 150 measured individuals, i.e. over the threshold;
    - or
    - 200, the original sample is used in the simulation given that the measured individuals in the subsample (150) is below the threshold.

In case of subsample differentiated by trip and commercial category, the minimum number by length accepted for subsampling is pertaining each category.

For the scenarios without sub-sampling the threshold should be set =1 and even the factor.

For scenarios of type 4, the user can choose to constraint the sub-sampling, setting a threshold higher than 1, or can decide to carry out the subsample, disregarding a check on the minimum number of individuals in the sample, just setting the threshold equal to 1 and the factor equal to the sub-sample factor (e.g. \( \frac{1}{2}, \frac{1}{4}, \ldots \)).

For testing different allocations of the trips among métier, the software provides 2 possibilities:

1. allow the bootstrap to extract randomly the samples, even if one métier is markedly more represented in the sampling data than another one;
2. force the bootstrap procedure to extract the realizations from the same number of samples for all the métier.

This can be implemented using the file recalled in the script #6 as the variable `path_table_trips_sharing`, where the number of trips to be extracted for each stratum is defined. In case the user would implement a random extraction at gear level putting together the different métier,
in the file recalled in the #6 script as path_table_NO_TRIPS_4_BOOTSTRAP, he/she should put the file NO_TRIPS_4_BOOTSTRAP (point 1 of the previous bullet point) equal to 0.

In the alternative case of unbalanced dataset, where for example one métier is more represented than another one, setting in the NO_TRIPS_4_BOOTSTRAP equal to the minimum number of trips available among the métier, the bootstrap extracts the samples with the same probability among the different métier (point 2 of the previous bullet point).

This feature is useful if the user needs to simulate a sampling design without up-weighting the métier more represented in the current sampling design. Analogously, can be tested a sampling design independently from the weight of the different quarters in the available data. This can be the case of years with not all the quarters covered by the sampling.

The possible re-allocation that can be simulated are:
- re-allocation of trips of level 4 OTB among OTB_DEF or OTB_MDD;
- re-allocation of annual trips among the quarters;
- re-allocation of trips of among the GSAs;
- re-allocation of trips of among different Countries.

The other specifications needed to run the script # 6 follow:
- SHORT_CASE_STUDY_NAME: the name of the simulated scenario
- target_SPECIES: the list of the target species with a short code (e.g. "DPS") and the scientific name reported in the CS and CL datasets (e.g. "Parapenaeus longirostris")
- nIter: the number of iterations to perform with the bootstrap
- THRESHOLD_TOLERANCE: a number used to define a tolerance around the thresholds (e.g. 0.01 that is 1%)

```r
# set the working directory
myWD <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19\SCRIPTS\SD Tool 2.01"
system.file_setup(myWD)

results_path <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19"

path_CL <- paste(results_path, "/input files/costCL.rds", sep="")
path_CS <- paste(results_path, "/input files/costCS.rds", sep="")

path_metier_shortcode <- paste(results_path, "/input files/metier_shortcode.csv", sep="")

# Set the absolute path of the file where the level of aggregation of the results is set
path_table_strat_results <- paste(results_path, "/input files/05_1_stratification_results.csv", sep="")

# Set the absolute path of the file where definition of the stratification of the sampling you want to reconstruct is set, e.g. random sampling in each stratum defined by GSA, year, level6
```
path_table_sampling_definition <<- paste(results_path, "/input files/06_1_sampling_design_definition.csv", sep="")

# Set the absolute path of the file where allocation of trips in the given sampling scheme is set
path_table_trips_sharing <<- paste(results_path, "/input files/06_2_sampling_design_trips_sharing.csv", sep="")

# Set the absolute path of the file where the subsample factors and the thresholds are defined by stratum
path_table_subsamples <<- paste(results_path, "/input files/06_3_subsample_definition.csv", sep="")

# Set the absolute path of the file where the minimum number of trips available for the bootstrap that can be equal for all the strata
path_table_NO_TRIPS_4_BOOTSTRAP <<- paste(results_path, "/input files/06_4_NO_TRIPS_4_BOOTSTRAP.csv", sep="")

SHORT_CASE_STUDY_NAME <<- "scenario_name"

target_SPECIES <<- list("DPS" = "Parapenaeus longirostris",
"ARA" = "Aristeus antennatus")

nIter <<- 100

THRESHOLD_TOLERANCE <<- 0.05

# ATTENTION!!! # BEFORE STARTING RUN THE CODE FIX THE BUG IN THE COST FUNCTION RUNNING THE FOLLOWING CODE:

fixInNamespace(spdAgreg, pos="package:COSTdbe")

# WHEN THE WINDOWS APPEARS, replace the line 8 with the following code:
# val <- paste(unlist(BY), collapse = "::")

• Open the **06_runScenario.Rmd** script with R studio and run clicking on knit button AT- TENTION: when the script is run, a window appears to allow the modification of a line of code in a COST function to fix a bug according the following indication in the code.
All the formats needed for running #6 script are stored in the input files folder, within the SD Tool directory. Namely:

- **05_1_stratification_results.csv** is the same produced by the data investigation and optimization phases;
- **06_1_sampling_design_definition.csv** (see Table B.15) defining strata as in the stratification_results.csv table. It could be considered as a list of the sampling strata in the considered domains, e.g. if the domain is GSA-year-OTB (lev4), the strata in the domain could be GSA-Y-OTB_DEF and GSA-Y-OTB_DWS (according to the metier/lev6) or, alternatively, GSA-1-OTB, GSA-2-OTB, GSA-3-OTB and GSA-4-OTB (according the quarters);
- **06_2_sampling_design_trips_sharing.csv** (see Table B.16) to set the number of trips for each stratum (column G). This number should be a compromise among the results obtained in the optimization process for each species. Columns A-B, C-D and E-F should be identical to each other and respectively also to the columns A, B, C in the output .csv “Optimized sample size”;
- **06_3_subsample_definition.csv** (see Table B.17) to simulate a subsample by trip/haul or by trip/haul and commercial category for each stratum and species (in case of no subsample set for each stratum and species factor ad threshold are = 1);
- **06_4_NO_TRIPS_4_BOOTSTRAP.csv** (see Table B.18) to inform the script if the extraction of the trips should be random among the strata or, in case of unbalanced sampling among the strata, it should be made applying the same probability of extraction of the samples in the different strata (see description above, in the same paragraph).

The outputs saved in .csv format in the folder 06_scenarioX (a different folder will be created for each scenario) are reported below (from Table B.19 to Table B.12).

Table B.15(*) - Structure of the 06_1_sampling_design_definition.csv table (it is the same of the stratification_results table) in input to the script #05_1 used to define the level of aggregation of the sampling in terms of spatial, temporal and technical unit.

<table>
<thead>
<tr>
<th>Space</th>
<th>time</th>
<th>Technical</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSA</td>
<td>Y</td>
<td>lev4</td>
</tr>
<tr>
<td>Country</td>
<td>Q</td>
<td>lev6</td>
</tr>
<tr>
<td>ALL</td>
<td>S</td>
<td>NONE</td>
</tr>
</tbody>
</table>

(*) All the results in output from script #6 will be returned according the level of aggregation defined in this table. Here only one of the items listed in the rows can be inserted to define the level of aggregation (e.g. GSA-Q-lev6).
Table B.16 – Structure of the 06_2_sampling_design_trips_sharing.csv table in input to the script #05_1.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong> space_original</td>
</tr>
<tr>
<td><strong>B</strong> Space</td>
</tr>
<tr>
<td><strong>C</strong> time_original</td>
</tr>
<tr>
<td><strong>D</strong> Time</td>
</tr>
<tr>
<td><strong>E</strong> technical_original</td>
</tr>
<tr>
<td><strong>F</strong> technical</td>
</tr>
<tr>
<td><strong>G</strong> no_of_trips</td>
</tr>
</tbody>
</table>

Table B.17 – Structure of the 06_3_subsample_definition.csv table containing the subsample factors and thresholds by stratum for each species and, if relevant, commercial category in input to the script #05_1. The strata should reflect the settings in the *stratification_results* table.

<table>
<thead>
<tr>
<th>species</th>
<th>space</th>
<th>Time</th>
<th>technical</th>
<th>commercial_category</th>
<th>factor</th>
<th>threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Parapenaeus longirostris</em></td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>1</td>
<td>0.125</td>
<td>170</td>
</tr>
<tr>
<td><em>Aristeus antennatus</em></td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>1</td>
<td>0.25</td>
<td>120</td>
</tr>
<tr>
<td><em>Parapenaeus longirostris</em></td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>2</td>
<td>0.125</td>
<td>90</td>
</tr>
</tbody>
</table>

Table B.18 – Structure of the 06_4_NO_TRIPS_4_BOOTSTRAP.csv table in input to the script #05_1 containing the number to be used for the bootstrap among the available trips for each stratum in order to re-construct the sampling using exactly the same number of trips for all the strata and avoid that the final CV is influenced by the availability of the trips in the different strata of the historical data. The strata should reflect the settings in the *stratification_results* table.

<table>
<thead>
<tr>
<th>Space</th>
<th>time</th>
<th>technical</th>
<th>min_accepted_ss</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>33</td>
</tr>
<tr>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>44</td>
</tr>
</tbody>
</table>
### Table B.19 – Structure of the table containing the values of the CVs by length from bootstrap in output from the script #06.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Species</strong></td>
</tr>
<tr>
<td><strong>Iteration</strong></td>
</tr>
<tr>
<td><strong>Var1</strong></td>
</tr>
<tr>
<td><strong>Var2</strong></td>
</tr>
<tr>
<td><strong>Var3</strong></td>
</tr>
<tr>
<td><strong>Length</strong></td>
</tr>
<tr>
<td><strong>CV</strong></td>
</tr>
</tbody>
</table>

- **Species**: Species
- **Iteration**: A number identifying the iteration
- **Var1**: Spatial stratification (e.g. GSA)
- **Var2**: Time stratification (e.g. 1, 2, 3, 4 for quarters; 1 for year)
- **Var3**: Metier/gear stratification
- **Length**: Length class
- **CV**: Coefficient of Variation by length class calculated according Vigneau & Mahévas (2007)

### Table B.20 – Structure of the table containing the values of the CVs from bootstrap in output from the script #06.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cv</strong></td>
</tr>
<tr>
<td><strong>nMeas</strong></td>
</tr>
<tr>
<td><strong>sampSize</strong></td>
</tr>
<tr>
<td><strong>Var1</strong></td>
</tr>
<tr>
<td><strong>Var2</strong></td>
</tr>
<tr>
<td><strong>Var3</strong></td>
</tr>
<tr>
<td><strong>Species</strong></td>
</tr>
</tbody>
</table>

- **Cv**: Coefficient of Variation calculated according Vigneau & Mahévas (2007)
- **nMeas**: Number of measured individuals
- **sampSize**: Number of trips
- **Var1**: Spatial stratification (e.g. GSA)
- **Var2**: Time stratification (e.g. 1, 2, 3, 4 for quarters; 1 for year)
- **Var3**: Metier/gear stratification
- **Species**: Species
Table B.21 – Structure of the table containing the rows of the HL table used in the bootstrap procedure in output from the script #05_1. The first 3 columns indicate the stratum in terms of spatial, temporal and technical stratification followed in the analyses and all the remaining columns follow the SDEF specifications. Depending on the performed scenario and on the settings related to the possibility to simulate a subsampling, the number of individuals in this table could be the same found in the original dataset or reduced according to the subsampling settings.

<table>
<thead>
<tr>
<th>Description</th>
<th>Var1</th>
<th>Var2</th>
<th>Var3</th>
<th>sampType</th>
<th>landCtry</th>
<th>vslFlgCtry</th>
<th>Year</th>
<th>Proj</th>
<th>trpCode</th>
<th>staNum</th>
<th>Spp</th>
<th>catchCat</th>
<th>landCat</th>
<th>commCatSc</th>
<th>subSampCat</th>
<th>Sex</th>
<th>lenCls</th>
<th>lenNum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial stratification (e.g. GSA)</td>
<td></td>
<td></td>
<td></td>
<td>“S” = sea sampling, “M” = market sampling of known fishing trips, “D” = market sampling of mixed trips, “V” = vendor.</td>
<td>ISO 3166–1 alpha-3 codes: the country where the vessel is landing and selling the catch</td>
<td>ISO 3166–1 alpha-3 codes. The flag country of the vessel. This may be different from the landing country</td>
<td></td>
<td></td>
<td>Trip ID code</td>
<td>Station number (haul)</td>
<td>Species</td>
<td>The fate of the catch: Dis = discard, Lan = landing</td>
<td>The intended usage at the time of landing. This should match the same field in the LS record (whether or not the fish was actually used for this or another purpose): IND = industry, HUC = human consumption</td>
<td>Commercial sorting scale code (optional for “Unsorted”)</td>
<td>Commercial sorting category in the given scale (optional for “Unsorted”). See (EC, 2006) and later amendments when scale is “EU”</td>
<td>Used when different fractions of the same species are subsampled at different levels. Typically used when few large specimens are removed from the total catch before the many small fish are subsampled</td>
<td>Male, = , Female, T = Transitional = (optional for “Unsexed”)</td>
<td>In mm. Identifier: lower bound of size class, e.g. 650 for 65 – 66 cm</td>
</tr>
</tbody>
</table>
Table B.22 – Structure of the table containing the LFD raised to the total landing resampled in the bootstrap procedure in output from the script #06.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
</tr>
<tr>
<td>Iteration</td>
</tr>
<tr>
<td>Var1</td>
</tr>
<tr>
<td>Var2</td>
</tr>
<tr>
<td>Var3</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Value</td>
</tr>
</tbody>
</table>

B4.4 Script Compare scenarios
- Open 07_runEvaluation.Rmd script in R studio and run it to compare the results from the different scenarios: all the results in the folders located in the case study path and starting with code “06” will be automatically read by the scripts and appended in a unique .csv file.
- The outputs in .csv format and the plots saved by the script in the folder 07_Evaluation are reported below.

Table B.23 – Structure of the table containing the CVs by length table by scenario in output from script #07.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario</td>
</tr>
<tr>
<td>Species</td>
</tr>
<tr>
<td>Iteration</td>
</tr>
<tr>
<td>Var1</td>
</tr>
<tr>
<td>Var2</td>
</tr>
<tr>
<td>Var3</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>CV</td>
</tr>
</tbody>
</table>
### Table B.24 – Structure of the table containing the CVs table by scenario in output from script #07.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario</strong></td>
</tr>
<tr>
<td><strong>Cv</strong></td>
</tr>
<tr>
<td><strong>nMeas</strong></td>
</tr>
<tr>
<td><strong>sampSize</strong></td>
</tr>
<tr>
<td><strong>Var1</strong></td>
</tr>
<tr>
<td><strong>Var2</strong></td>
</tr>
<tr>
<td><strong>Var3</strong></td>
</tr>
<tr>
<td><strong>Species</strong></td>
</tr>
<tr>
<td><strong>Iteration</strong></td>
</tr>
</tbody>
</table>

### Table B.25 – Structure of the table containing the LFD raised to the total landings from bootstrap by scenario in output from script #07.

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario</strong></td>
</tr>
<tr>
<td><strong>Species</strong></td>
</tr>
<tr>
<td><strong>Iteration</strong></td>
</tr>
<tr>
<td><strong>Var1</strong></td>
</tr>
<tr>
<td><strong>Var2</strong></td>
</tr>
<tr>
<td><strong>Var3</strong></td>
</tr>
<tr>
<td><strong>Length</strong></td>
</tr>
<tr>
<td><strong>Value</strong></td>
</tr>
</tbody>
</table>
Table B.26 – Example of the summary table saved by script #7. For each stratum and species, the table contains the values of CV (%) calculated according to Vigneau & Mahevas (2007), the sample size (in terms of numbers of trips) and the number of measured individuals simulated in each scenario (Baseline included). Var1, Var2 and Var3 indicate respectively the space, time and technical stratification.

<table>
<thead>
<tr>
<th>species</th>
<th>Var1</th>
<th>Var2</th>
<th>Var3</th>
<th>Scenario</th>
<th>CV</th>
<th>samp_size</th>
<th>no_indiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parapenaeus longirostris</td>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>Baseline</td>
<td>14.14</td>
<td>42</td>
<td>49359</td>
</tr>
<tr>
<td>Parapenaeus longirostris</td>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>Scenario1</td>
<td>19.72</td>
<td>19</td>
<td>22179</td>
</tr>
<tr>
<td>Parapenaeus longirostris</td>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2A</td>
<td>30.57</td>
<td>42</td>
<td>21415</td>
</tr>
<tr>
<td>Parapenaeus longirostris</td>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2B</td>
<td>46.84</td>
<td>42</td>
<td>12880</td>
</tr>
<tr>
<td>Aristeus antennatus</td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>Baseline</td>
<td>18.06</td>
<td>35</td>
<td>9505</td>
</tr>
<tr>
<td>Aristeus antennatus</td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>Scenario1</td>
<td>19.14</td>
<td>32</td>
<td>9009</td>
</tr>
<tr>
<td>Aristeus antennatus</td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2A</td>
<td>42.55</td>
<td>35</td>
<td>5418</td>
</tr>
<tr>
<td>Aristeus antennatus</td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2B</td>
<td>57.29</td>
<td>35</td>
<td>4002</td>
</tr>
</tbody>
</table>

Table B.27 – Example of the summary table saved by script #7. It contains for each stratum and species the change in percentage of the CV, the sample size (in terms of numbers of trips) and the number of measured individuals simulated in each scenario compared to the Baseline scenario (percentage change). The value of the Earth Mover Distance (EMD) estimated between the raised length distribution simulated in the scenario and that of the Baseline scenario is also reported. Var1, Var2 and Var3 indicate respectively the space, time and technical stratification.

<table>
<thead>
<tr>
<th>species</th>
<th>Var1</th>
<th>Var2</th>
<th>Var3</th>
<th>Scenario</th>
<th>Change in CV</th>
<th>Change in samp_size</th>
<th>Change in no_indiv</th>
<th>EMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parapenaeus longirostris</td>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>Scenario1</td>
<td>39.45</td>
<td>-54.76</td>
<td>-55.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Parapenaeus longirostris</td>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2A</td>
<td>116.17</td>
<td>0</td>
<td>-56.61</td>
<td>0.39</td>
</tr>
<tr>
<td>Parapenaeus longirostris</td>
<td>GSA18</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2B</td>
<td>231.28</td>
<td>0</td>
<td>-73.91</td>
<td>0.58</td>
</tr>
<tr>
<td>Aristeus antennatus</td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>Scenario1</td>
<td>5.98</td>
<td>-8.57</td>
<td>-5.21</td>
<td>0.21</td>
</tr>
<tr>
<td>Aristeus antennatus</td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2A</td>
<td>135.58</td>
<td>0</td>
<td>-43</td>
<td>1.21</td>
</tr>
<tr>
<td>Aristeus antennatus</td>
<td>GSA19</td>
<td>1</td>
<td>OTB</td>
<td>Scenario2B</td>
<td>217.21</td>
<td>0</td>
<td>-57.9</td>
<td>1.65</td>
</tr>
</tbody>
</table>
Figure B.2 – Example of bubble plot with the CV values obtained in the performed scenarios (number of trips on the x-axis and number of measured individuals on the y-axis).

Figure B.3 – Example of the bubble plot of the EMD values in the performed scenarios.
Figure B.4 – Example of plot of the length frequency distributions (LFD) raised to the total landings obtained in the performed scenarios. The same output is produced by trip and commercial category.

Figure B.5– Example of plot with the reduction of the numbers of measured individuals (y-axis) by trip/haul (x-axis). Scenario compared to the HL table (original without reduction) used in the baseline.
B5. Case study example

Several applications of the SDTool can be found in the STREAM deliverable 3.3.

An application of the SDTool was carried out during the WKBIOPTIM3 on the commercial data of *M. barbatus* in Southern Adriatic Sea. The results were compared with the ones obtained with FishPi4WKBIOPTIM script. For details, see Section 5.2 and Section 5.5.

B6. Conclusions

The new version of SDTool, improved and generalized within STREAM project, allows to explore different sampling designs, evaluating the impact of different stratifications on the CV and on the EMD.

Despite on the flexibility of the tool, allowing to test different spatial, temporal and technical stratifications, SDTool is based on bootstrap, thus the use of it with limited number of observations could return unreliable results to be considered with caution.

B7. Improvements required

The inclusion of more indicators would undoubtedly be beneficial to provide additional information on the impact of alternative sampling designs on the LFDs.

The development of an R tool including all the scripts of SDTool would be also important to allow a wider availability to the method and to ease the application to other case studies.
B8. GitHub Link


https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SD%20Tool%202.04

B9. References


C1. Introduction

BioSim is based on the methodology developed at ICES WKBIOPTIM (ICES WKBIOPTIM Final Report, 2017) to identify the “optimal” number of individuals to be measured “at sample level” (i.e. per trip/haul).

In the STREAM project (Strengthening Regional cooperation in the area of fisheries biological data collection in the Mediterranean and Black Sea), the original scripts were updated including the possibility to evaluate the impact on the biological estimates (maturity, sex ratio at length, etc.) globally rather than by trip, allowing to focus on the final outcome of the sampling related to biological variables.

From the analyses on the length distributions by trip/haul the minimum number of fish to be measured can be found, thus allowing to obtain representative samples of the sampled population, and avoid oversampling (e.g. several hundreds of measurements per sample, over a relatively short size/age-class range). This “optimal” number of individuals to be measured can be used as a guide for exploring the extent of different subsamplings in the definition of the scenario with SDTool.

C2. Libraries required

The following libraries are needed to run the scripts: rlist, dplyr, parallel, ggplot2, reshape2, emdist, reshape, knitr, pander, data.table.

C3. Input data

The data to be used to feed BioSim are the sampling data by primary unit (e.g. trip/haul), by species, commercial category (not mandatory) and length.

C3.1 Dataset format

The BioSim Tool works on individual measures taken from the COST CA table. It uses the COST HH table to associate the information on the individuals to the haul by métier/gear and quarter.

C3.2 Description of the variables of the dataset

For the description of the variables stored according the COST format, refer to Jansen (2009).

C3.3 Initial values/starting values for simulations

A dummy dataset is available on github at: https://github.com/aleligas/STREAM_MARE-2016-22/tree/Task-3.3---Sampling-Strategy-Optimization.

C4 Scripts files

The BioSim Tool consists of the following 5 main scripts (see also Figure C.7):

- A_data_preparation.R for the preparation of the dataset starting from CA and HH tables. An .RData file is saved in the Input files folder and used by the next scripts.
• **B_data_simulation_LENGTH.R** performs simulations, resampling individual length measurements from the original samples for a given number of iterations and for different sample size (scenarios) and saves indicators (MWCV, n_class_sampled, n_modes_correct) to be compared to the current ones.

• **C_data_simulation_MATURITY.R** carry out simulations, resampling individual maturity measurements from the original sampling for a given number of iterations and for different sample size (scenarios) and saves the proportion of mature individuals at length, the CV and the EMD (Earth Mover Distance) to be compared with the current ones.

• **D_data_simulation_SEX-RATIO.R** performs simulations, resampling individual sex measurements from the original samples for a given number of iterations and for different sample size (scenarios) and saves the sex-ratio at length, the CV and the EDM (Earth Mover Distance) to be compared with the original sample.

• **E_data_simulation_AGE.R** performs simulations, resampling individual age measurements according to different sample size by length class (scenarios) from the original sampling for a given number of iterations and saves the age structure, the CV and the EDM (Earth Mover Distance) to be compared with the original sample.

A script named funs.r contains auxiliary functions for the running of the main scripts.

![Figure C.7 – Configuration of the main folder of the BioSim Tool 1.01 with main scripts and auxiliary functions.](image-url)
C4.1 Script Data preparation (script #A)

- Open the script `A_data_preparation.R` with R or R studio and compile the settings indicating the following elements:
  - `caseStudy_path`: the path of the directory of the case study, where the input files are located and all the outputs will be saved in the related folders (created by the tool);
  - `dir_res`: the path of the directory where the output will be saved, i.e. `A_data_preparation`;
  - `df`: the path where the SDEF CA table is located;
  - `hh`: the path where the SDEF HH table is located;
  - `IMMATSTAGES`: the list of the maturity stages considered immatures, differentiated by the maturity scale used in the dataset (for each maturity scale a vector should be inserted, e.g. `c("Medits scale", "0","1","2a")`);
  - `ref_tab`: the default table provided with the code used for the conversion of the columns names;
  - `metier_shortcode`: the path where the table of correspondences between Lev6 and Lev4 is located.

```r
# load file
df=read.table(paste(caseStudy_path, "\input files\SDEF CS-CA data.csv", sep=""), sep=";", header=T)

# Inclusion of date and metier
hh=read.table(paste(caseStudy_path, "\input files\SDEF CS-HH data.csv", sep=""), sep=";", header=T)

# selection of immature stages made by the user
IMMATSTAGES = list(c("0","1","2a"))

# formatting columns names
ref_tab<-read.csv2(paste(caseStudy_path, "\input files\A_col_names_conversion_table.csv", sep=""), sep=";", header=T)

# association of lev4 to each trip.
metier_shortcode=read.table(paste(caseStudy_path, "\input files\metier_shortcode.csv", sep=""), sep=";", header=T)
```
- Run script #A to prepare the data for the analysis on the optimization of the biological variables;
- The output R object is saved by the script in the folder A_data_preparation.

### C4.2 Script Optimize sampling size (number of length measurements) (script #B)

- Compile the settings in the script `B_data_simulation_LENGTH.R` indicating the following elements:
  - `myWD`: the path of the working directory, i.e. the path where BioSim Tool folder is located;
  - `caseStudy_path`: the path of the directory of the case study, where the input files are located and all the outputs will be saved in the related folders (created by the tool);
  - `species_name`: the name of the species used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - `short_name`: a short code for the species analysed to be used for the output files names;
  - `GSA`: the area used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - `by_cat`: “Y” or “N” indicating if the analyses have to be done considering as sample the trip (if N) or the trip and commercial category (if Y);
  - `by_EUlev6`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the metier/lev6 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - `by_EUlev4`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the gear/lev4 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - `by_quart`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the quarters or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - `variable_table`: the table where the variable (length class) and its type (numerical) are indicated. In this table additional inputs are required in order to perform an investigation on the samples available in the dataset: for each sample the length distribution in the dataset is compared with a new length distribution built with a wider length class span. This is done with the aim of verifying if using a different precision in the length measurement (e.g. for Merluccius merluccius a precision equal to 1 cm instead of 0.5 cm) could lead to significant changes in the length distribution. Given that sometimes the distribution can present some peaks that are identified as modes by chance (false modes), a minimum proportion of individuals respect to the total number in the sample is required to be used as threshold to accept a mode as a real mode. The other elements required to perform the described analysis are: the original class span (in mm, e.g. 0.5) and the smoothed class span (in mm, e.g. 1) (see Table C.28).
Table C.28– The table on the characteristics of the input variables to the B_data_simulation_LENGTH.R script.

<table>
<thead>
<tr>
<th>variable</th>
<th>type</th>
<th>Original class_span</th>
<th>Smooth class_span</th>
<th>min_proportion to_accept_mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>lenCls</td>
<td>numerical</td>
<td>1</td>
<td>2</td>
<td>0.01</td>
</tr>
</tbody>
</table>

- n_sims: number of iterations to be run;
- samp_sizes: a vector containing the sample size to be simulated e.g. c(30, 40, ..., 200) by sample;
- sampling_design, a list containing the following elements:
  - stratified: TRUE or FALSE indicates if the sampling is stratified or not. In script #B it is always FALSE;
  - strata_var: the variable for the stratification. In script #B it is always an empty string because the stratified field is FALSE;
- sampling_options, a list containing the following elements:
  - stages: indicates if the sampling is in one or two stages. In the present version of the script only the “one” stage has been tested;
  - stratified: TRUE or FALSE indicates if the sampling is stratified (by length class) or not. In script #B (length measures optimization) it is always FALSE;
  - strata_var: the variable for the stratification. In script #B it is always an empty string, because the stratified field is FALSE;
  - stage1_samp_size: default value NA;
  - replacement: indicate if the bootstrap is done with or without replacement. In script #B the bootstrap is always without replacement.
  - sample_all_available: default value TRUE;
  - sample_all_available_warning: default value TRUE
  - vars_to_keep: filled in only when the sampling is stratified (by length class). In script #B it is always an empty string.
- SAVE_PLOTS_BY_TRIP: TRUE or FALSE indicates if you want to save the plots of the simulation for each trip.
- load(paste(caseStudy_path, "\input files\input_data.rdata", sep=""): the path where the input_data.rdata is located.

```r
# set the working directory
myWD <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19\\SCRIPTS\BioSim Tool 1.01"
setwd(myWD)

caseStudy_path <-"E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19"

species_name <- "Parapenaeus longirostris"
short_name <- "DPS"
```
GSA <- "GSA19" # "ALL"

# stratification of sampleId definition
by_cat="N" # "Y" or "N"
by_EUlev6="N" # "Y" or "N"
by_EUlev4="N"
by_quart="N" # "Y" or "N"

# read variable table
variable_table <- read.csv2(paste(caseStudy_path, "\input files\B_variable_table.csv", sep=""), as.is=TRUE)

# expliciting the number of iterations and the sample sizes
n_sims <- 50
samp_sizes <- c(seq(30, 200, by=10))
# e.g. between the intermediate quantiles

# set sampling design of sample data
sampling_design <- list(stratified = FALSE, strata_var = "")

sampling_options <- list(n_sims = n_sims,
                         stages="one", # no of stages
                         stratified = FALSE,
                         strata_var = "", # stratification details
                         stage1_samp_size=NA,
                         samp_sizes = samp_sizes, # samp sizes
                         replacement=FALSE,
                         sample_all_available = TRUE,
                         sample_all_available_warning = TRUE, # replacement options
                         vars_to_keep = c(""))

SAVE_PLOTS_BY_TRIP <- FALSE

# load input data
load(paste(caseStudy_path, "\input files\input_data.rdata", sep=""))
For the choice of the steps to be simulated for the sample sizes, it is suggested to take as minimum sample size a number of individuals that would cover all the length classes generally present in the LFD of the species. For example, for a deep water rose shrimp the lengths of caught individuals range from about 5 to 35 mm, then an LFD should contain about 30 length classes of 1 mm. The minimum number of individuals to be considered could be, for example, 30 by trip/haul.

After the compilation of the settings, run the script #B until the row 145. The summary table with the quantiles of the numbers of individuals by trip/haul will be saved in the output folder.

The summary table with the quantiles can support the selection of the most representative trips to be used for the bootstrap (see code below) by setting a minimum number of individuals by trip/haul (or by trip and commercial category) as threshold, and consequently the choice of the sample sizes to be simulated.

Compile the settings in the script B_data_simulation_LENGTH.R indicating the `min_n` to be used for the selection of the most representative trips.

```r
# setting of the minimum number of individuals considered representative
min_n <- 50
```

After setting `min_n` at row 145, run again all the script #B to obtain the values of indicators for each sample size.

The outputs tables in .csv format and the plots saved by the script in the folder B_length are reported below.

**Table C.29** – Example of the table saved by the script #B, the table reports the annual numbers of individuals measured in the past years and available in the input dataset (from the historical data). In the last row the average calculated over the years is reported.

<table>
<thead>
<tr>
<th>Metier_Gear</th>
<th>quarter</th>
<th>commCat</th>
<th>Year</th>
<th>length_measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014</td>
<td>43243</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2015</td>
<td>58792</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2016</td>
<td>79189</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2017</td>
<td>53738</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014-2017</td>
<td>58740</td>
</tr>
</tbody>
</table>
Table C.30 – Example of the table saved by the script #B. The table reports the values of EMD and mean weighted CV (%) calculated as average on the iterations of the total LFD (summed up on all the considered trips). The last line indicates the value of the MWCV% calculated on the original sample (baseline) and the total number of the measured individuals is reported in brackets.

<table>
<thead>
<tr>
<th>commCat</th>
<th>Metier_Gear</th>
<th>quarter</th>
<th>samp_size</th>
<th>EMD</th>
<th>MWCV %</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>30</td>
<td>0.435</td>
<td>6.5</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>40</td>
<td>0.432</td>
<td>5.6</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>50</td>
<td>0.435</td>
<td>5</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>60</td>
<td>0.43</td>
<td>4.6</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>70</td>
<td>0.436</td>
<td>4.3</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>80</td>
<td>0.432</td>
<td>4</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>90</td>
<td>0.433</td>
<td>3.8</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>100</td>
<td>0.436</td>
<td>3.6</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>110</td>
<td>0.428</td>
<td>3.4</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>120</td>
<td>0.438</td>
<td>3.3</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>130</td>
<td>0.431</td>
<td>3.1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>140</td>
<td>0.434</td>
<td>3</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>150</td>
<td>0.433</td>
<td>2.9</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>160</td>
<td>0.435</td>
<td>2.8</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>170</td>
<td>0.431</td>
<td>2.7</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>180</td>
<td>0.434</td>
<td>2.7</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>190</td>
<td>0.432</td>
<td>2.6</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>200</td>
<td>0.438</td>
<td>2.5</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Baseline [1382]</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure C.8 – Example of the plot of the EMD calculated as average on the iterations of the total LFD (summed up on all the considered trips).

Figure C.9 – Example of the plot of the MWCV\% (y-axis) calculated as average on the iterations of the total LFD (summed up on all the considered trips) reported for each simulated sample size (x-axis). The red line indicates the value of the MWCV\% calculated on the original sample (baseline) with the total number of measured individuals. The dotted blue lines represent the percentiles (5th, 25th, median, 75th and 95th) estimated on the overall simulations.

The MWCV\% plot should be considered together with the EMD plot. An “optimal” sample size in terms of number of individuals can be the number that can determine a value of MWCV\% sufficiently low and stable e.g. below the 25-th percentiles or below the median value of all the MWCV\% values. Considering that the MWCV\%, differently from the Vigneau & Mahevas (2007) formula, does not take into account the number of trips in the stratum and the commercial landings, it is convenient also to look at the EMD and make different hypothesis of optimal sample
size to be tested in the scenarios (with SDTool) through the calculation of the Vigneau & Mahevas (2007) CV. A good compromise could be the number from which the EMD starts to become more stable increasing the sample size.

- After the identification of the better performing sample size(s), compile the script `B2_calculate_subsample.R` indicating the following elements, as in the script #B:
  - `myWD`: the path of the working directory, i.e. the path where BioSim Tool folder is located;
  - `caseStudy_path`: the path of the directory of the case study, where the input files are located and all the outputs will be saved in the related folders (created by the tool);
  - `species_name`: the name of the species used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - `short_name`: a short code for the analysed species to be used for the output files names;
  - `GSA`: the area used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - `by_cat`: “Y” or “N” indicating if the analyses have to be done considering as sample the trip (if N) or the trip and commercial category (if Y);
  - `by_EUlev6`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the metier/lev6 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - `by_EUlev4`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the gear/lev4 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - `by_quart`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the quarters or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - `load(paste(caseStudy_path, \\input files\\input_data.rdata", sep=""))`: the path where the `input_data.rdata` is located must be set in the code
  - `thresholds_name`: a name for the output file
  - `thresholds_table`: the path where the thresholds table (see Table C.31) is located must be set in the code.

```r
# set the working directory
myWD <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19\SCRIPTS\BioSim Tool 1.01"
setwd(myWD)

caseStudy_path <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19"

species_name <- "Aristeus antennatus"
short_name <- "ARA"
```
GSA <- "GSA19" # "ALL"

# stratification of sampleId definition
by_cat="N" # "Y" or "N"
by_EUlev6="N" # "Y" or "N"
by_EUlev4="N"
by_quart="N" # "Y" or "N"

# load input data
load(paste(caseStudy_path, "\input files\input_data.rdata", sep=""))

# SET THE CHOSEN THRESHOLD TO ESTIMATE THE RESPECTIVE SUBSAMPLE FACTOR:
# for the name for the output file
thresholds_name <- "catN_lev4_quartY"
thresholds_table <- read.csv2(paste(caseStudy_path, "\input files\HKE_thresholds_catN_lev4_quartY.csv", sep=""), as.is=TRUE)

Table C.31 – Example of the input table with the thresholds chosen as optimal numbers by commercial category, gear/metier and quarter. The table has the same structure of the table in output from the script #B and can be extrapolated from that table.

<table>
<thead>
<tr>
<th>commCat</th>
<th>Metier_Gear</th>
<th>quarter</th>
<th>samp_size</th>
<th>EMD</th>
<th>MWCV %</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>OTB</td>
<td>2</td>
<td>40</td>
<td>0.69</td>
<td>15.6</td>
</tr>
<tr>
<td>-1</td>
<td>OTB</td>
<td>3</td>
<td>35</td>
<td>0.367</td>
<td>17.1</td>
</tr>
<tr>
<td>-1</td>
<td>OTB</td>
<td>4</td>
<td>45</td>
<td>0.332</td>
<td>9.4</td>
</tr>
</tbody>
</table>

• The outputs tables in .csv format saved by the script in the folder B_length are reported in Table C.32 and Table C.33.
Table C.32 – Example of the table with subsample factors (by commercial category, gear/metier and quarter), calculated averaging the ratios defined for each trip as the numbers of lengths in the original dataset to the “optimal” number of lengths suggested from the simulations (only the ratio greater than 1 have been considered in the average). In the columns Nb_over_Thr and Tot the number of the samples with a number of lengths greater than the “optimal” number (threshold) and the total number of samples in the original dataset are respectively reported.

<table>
<thead>
<tr>
<th>commCat</th>
<th>Metier_Gear</th>
<th>quarter</th>
<th>Sub_sample factor</th>
<th>Nb_over_Thr</th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>OTB</td>
<td>2</td>
<td>2</td>
<td>24</td>
<td>58</td>
</tr>
<tr>
<td>-1</td>
<td>OTB</td>
<td>3</td>
<td>2</td>
<td>21</td>
<td>37</td>
</tr>
<tr>
<td>-1</td>
<td>OTB</td>
<td>4</td>
<td>2</td>
<td>45</td>
<td>80</td>
</tr>
</tbody>
</table>

Table C.33 – Example of the table containing all the samples from the dataset classified by commercial category, gear/metier and quarter with indication of number of individuals, the threshold and the ratio between the original number of measured individuals and the threshold.

<table>
<thead>
<tr>
<th>sampId</th>
<th>Num_ind</th>
<th>quarter</th>
<th>commCat</th>
<th>Metier_Gear</th>
<th>thr</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_ITA_1</td>
<td>54</td>
<td>3</td>
<td>-1</td>
<td>OTB</td>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>105_ITA_1</td>
<td>50</td>
<td>3</td>
<td>-1</td>
<td>OTB</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>106_ITA_1</td>
<td>39</td>
<td>3</td>
<td>-1</td>
<td>OTB</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>111_ITA_1</td>
<td>50</td>
<td>3</td>
<td>-1</td>
<td>OTB</td>
<td>35</td>
<td>1</td>
</tr>
</tbody>
</table>

C4.3 Script Optimize number of maturity measurements (script #C)

- Open the `C_data_simulation_MATURITY.R` script in R or R studio and compile the settings in the script indicating the following elements:
  - `myWD`: the path of the working directory, i.e. the path where BioSim Tool folder is located;
  - `caseStudy_path`: the path of the directory of the case study where the input files are located and all the outputs will be saved in the related folders;
  - `species_name`: the name of the species used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - `short_name`: a short code for the analysed species to be used for the output files names;
  - `GSA`: the area used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - `SEX`: “M” or “F” or “ALL” indicating if the analyses have to be done considering respectively only males, only females or both;
  - `by_cat`: “Y” or “N” indicating if the analyses have to be done considering as sample the trip (if N) or the trip and commercial category (if Y);
  - `by_EUlev6`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the metier/lev6 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
- `by_EUlev4`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the gear/lev4 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
- `by_quart`: “Y” or “N” indicating if the results of the analyses have to be grouped considering the quarters or not (notice that this aggregation is used only to summarize the results and not in the analysis);
- `n_sims`: number of iterations to be run
- `samp_sizes`: a vector containing the sample size to be simulated e.g. c(30, 40, …, 200) by sample;
- `sampling_design`, a list containing the following elements:
  - `stratified`: TRUE or FALSE indicates if the sampling is stratified (by length class) or not. In the script #C it is always FALSE;
  - `strata_var`: the variable for the stratification. In script #C it is always an empty string, because the stratified field is FALSE;
- `sampling_options`, a list containing the following elements:
  - `stages`: indicates if the sampling is at one stage or two stages. In the present version of the script only the “one” stage has been tested;
  - `stratified`: TRUE or FALSE indicates if the sampling is stratified or not. In script #C it is always FALSE;
  - `strata_var`: the variable for the stratification. In script #C it is always an empty string, because the stratified field is FALSE
  - `stage1_samp_size`: default value NA;
  - `replacement`: indicate if the bootstrap is done with or without replacement. In script #C the bootstrap is always without replacement;
  - `sample_all_available`: default value TRUE;
  - `sample_all_available_warning`: default value TRUE;
  - `vars_to_keep`: filled in only when the sampling is stratified. In script #C it is always an empty string, because the stratified field is FALSE;
- `load(paste(caseStudy_path, "\input files\input_data.rdata", sep=""))`: the path where the `input_data.rdata` is located.

```r
# set the working directory

myWD <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19\SCRIPTS\BioSim Tool 1.01"
setwd(myWD)

caseStudy_path <-"E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19"

species_name <- "Aristeus antennatus"
short_name <- "ARA"

GSA <- "GSA19" # "ALL"
```
After the compilation of the settings, run the script #C until the row 162. The summary table reported in the R console with the quantiles of the numbers of individuals by trip/haul will be saved in the output folder.
The summary table with the quantiles can support the selection of the sample size that the trips to be involved in the resampling should have to ensure more reliable results.

- Compile the last setting in the script C_data_simulation_MATURITY.R indicating the min_n to be used for the selection of the most representative trips. In the case of simulations in a selected sample size (e.g. 10,20,…,200), please pay attention to select as min_n equal to the maximum of the range (e.g. 200) to allow the script to carry out the resampling without replacement.

```r
# setting of the minimum number of individuals considered representative
min_n <- 50
```

- After setting min_n at row 162, run again all the script #C to obtain the values of indicators for each sample size.
- The outputs tables in .csv format and the plots saved by the script in the folder C_maturity are below reported.

Table C.32 – Example of the table saved by the script #C, the table reports the annual numbers of individuals for which the maturity stage has been registered in the past years in the available input dataset (from the historical data). In the last row the average calculated over the years is reported.

<table>
<thead>
<tr>
<th>Metier_Gear</th>
<th>quarter</th>
<th>commCat</th>
<th>year</th>
<th>maturity_measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014</td>
<td>22915</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2015</td>
<td>32604</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2016</td>
<td>39225</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2017</td>
<td>28841</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014-2017</td>
<td>30896</td>
</tr>
</tbody>
</table>
Table C.33 – Example of the table saved by the script #C. The table reports the values of the EMD and mean weighted CV (%) calculated as average on the iterations of the total proportion of mature individuals by length class (summed up on all the considered trips). The last line indicates the value of the MWCV% calculated on the original sample (baseline) and the total number of the individuals for which the maturity stage was recorded (in parenthesis). The sample size indicated for the baseline is calculated as average of the individuals with maturity stage in the samples of the dataset.

<table>
<thead>
<tr>
<th>Metier_Gear</th>
<th>quarter</th>
<th>commCat</th>
<th>samp_size</th>
<th>CV %</th>
<th>EMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>15</td>
<td>1.67</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>25</td>
<td>1.29</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>35</td>
<td>1.10</td>
<td>0.001</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>45</td>
<td>0.97</td>
<td>0.001</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>55</td>
<td>0.87</td>
<td>0.001</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>65</td>
<td>0.81</td>
<td>0.001</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Baseline [727]</td>
<td>0.24</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure C.10 – Example of the plot of maturity at length calculated as average on the iterations of the total proportion of mature individuals by length class (summed up on all the considered trips). In each rectangle a different hypothesis of sample size is done, i.e. a different number of sex measures is considered by trip/haul. The blue points represent the simulated values and the red points the observed ones in the original sampling.
Figure C.11 - Example of the plot of the MWCV % (y-axis) calculated as average on the iterations of the proportion of mature individuals by length class (summed up on all the considered trips) considering different sample sizes (x-axis) by trip/haul. The red line indicates the value of the MWCV% calculated on the original sample (baseline) and the total number of the individuals for which the maturity stage was recorded. The dotted blue lines represent the percentiles (5th, 25th, median, 75th and 95th) estimated on the overall simulations.

The MWCV% plot should be considered together with the plot of maturity at length. An “optimal” sample size in terms of number of individuals for which maturity should be recorded can be the number corresponding to a MWCV% sufficiently low and stable e.g. below the 25-th percentiles or below the median value of all the MWCV% values (Figure C.11). It is convenient also to look at the plot of the maturity at length (Figure C.10) with different hypothesis of sample size in each rectangle.

C4.3 Script Optimize number of sex measurements (script #D)

- Open the script D_data_simulation_SEX-RATIO.R in R or R studio and compile the settings indicating the following elements:
  - myWD: the path of the working directory, i.e. the path where BioSim Tool folder is located;
  - caseStudy_path: the path of the directory of the case study where the input files are located and all the outputs will be saved in the related folders (created by the tool);
  - species_name: the name of the species used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - short_name: a short code for the analysed species to be used for the output files names;
  - GSA: the area used for the selection of the data from the case study dataset (including all the target species and GSAs);
- by_cat: “Y” or “N” indicating if the analyses have to be done considering as sample the trip (if N) or the trip and commercial category (if Y);
- by_EUlev6: “Y” or “N” indicating if the results of the analyses have to be grouped considering the metier/lev6 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
- by_EUlev4: “Y” or “N” indicating if the results of the analyses have to be grouped considering the gear/lev4 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
- by_quart: “Y” or “N” indicating if the results of the analyses have to be grouped considering the quarters or not (notice that this aggregation is used only to summarize the results and not in the analysis);
- n_sims: number of iterations to be run;
- samp_sizes: a vector containing the sample size to be simulated e.g. c(30, 40, ..., 200) by sample;
- sampling_design, a list containing the following elements:
  - stratified: TRUE or FALSE indicates if the sampling is stratified (by length class) or not. In script #D it is always FALSE;
  - strata_var: the variable for the stratification. In script #D it is always an empty string, because the stratified field is FALSE;
- sampling_options, a list containing the following elements:
  - stages: indicates if the sampling is at one stage or two stages. In the present version of the script only the “one” stage has been tested;
  - stratified: TRUE or FALSE indicates if the sampling is stratified or not. In script #D it is always FALSE;
  - strata_var: the variable for the stratification. In script #D it is always an empty string, because the stratified field is FALSE
  - stage1_samp_size: default value NA;
  - replacement: indicate if the bootstrap is done with or without replacement. In script #D the bootstrap is always without replacement.
  - sample_all_available: default value TRUE;
  - sample_all_available_warning: default value TRUE;
  - vars_to_keep: filled in only when the sampling is stratified. In script #D it is always an empty string, because the stratified field is FALSE;
- load(paste(caseStudy_path, “\input files\input_data.rdata", sep="")): the path where the input_data.rdata is located.

```
# set the working directory
myWD <- "E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19\SCRIPTS USATI\BioSim Tool 1.01"
setwd(myWD)

caseStudy_path <-"E:\11_STREAM\WP3\CASE STUDIES\1st_CS6_DPS_ARA_GSA18_19"

species_name <-"Parapenaeus longirostris"
```
short_name <- "DPS"
GSA <- "GSA18" # "ALL"

# set sampling design of sample data
sampling_design <- list(stratified = FALSE, strata_var = "")

# expliciting the number of iterations and the sample sizes
n_sims <- 50
samp_sizes <- c(seq(30, 170, by=10))

# stratification of sampleId definition
by_cat="N" # "Y" or "N"
by_EUlev6="N" # "Y" or "N"
by_EUlev4="N"
by_quart="N" # "Y" or "N"

# Sampling of different number of individuals without replacement
sampling_options <- list(n_sims = n_sims,
                          stages="one", # no of stages
                          stratified = FALSE,
                          strata_var = "", # stratification details
                          stage1_samp_size=NA,
                          samp_sizes = c(samp_sizes), # samp sizes
                          replacement=FALSE,
                          sample_all_available = TRUE,
                          sample_all_available_warning = TRUE, # replacement options
                          vars_to_keep = c(""))

# load input data
load(paste(caseStudy_path, "\input files\input_data.rdata", sep=""))
...
After the compilation of the settings, run the script #D until the row 162. The summary table with the quantiles of the numbers of individuals by trip/haul will be saved in the output folder.

The summary table with the quantiles can support the selection of the sample size that the trips to be involved in the resampling should have to ensure more reliable results.

Compile the last setting in the script D_data_simulation_SEX-RATIO.R indicating the min_n to be used for the selection of the most representative trips. In the case of simulations in a selected sample size (e.g. 10, 20…, 200), please pay attention to select as min_n equal to the maximum of the range (e.g. 200) to allow the script to carry out the resampling without replacement.

```r
# setting of the minimum number of individuals considered representative
min_n <- 50
```

After setting min_n at row 162, run again all the script #D to obtain the values of indicators for each sample size.

The outputs tables in .csv format and the plots saved by the script in the folder D_sexratio are below reported.

### Table C.34– Example of the table saved by the script #D. The table reports the annual numbers of individuals for which the sex has been recorded in the past years (from the historical data). In the last row the average calculated over the reported years.

<table>
<thead>
<tr>
<th>Metier_Gear</th>
<th>quarter</th>
<th>commCat</th>
<th>year</th>
<th>sex_measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014</td>
<td>38665</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2015</td>
<td>53061</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2016</td>
<td>69774</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2017</td>
<td>48061</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014-2017</td>
<td>52390</td>
</tr>
</tbody>
</table>
Table C.35 – Example of the table saved by the script #D. The table reports the values of EMD and the MWCV% calculated as average on the iterations of the total sex-ratio (females/(males+females)) by length class (summed up on all the considered trips). The last line indicates the value of the MWCV% calculated on the original sample (baseline) and the number of the individuals for which the sex was recorded for each trip/haul (in brackets). The sample size indicated for the baseline is calculated as average of the number of sexed individuals of the samples in the dataset.

<table>
<thead>
<tr>
<th>Metier_Gear</th>
<th>quarter</th>
<th>commCat</th>
<th>samp_size</th>
<th>CV (%)</th>
<th>EMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>30</td>
<td>4.51</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>40</td>
<td>3.93</td>
<td>0.05</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>50</td>
<td>3.54</td>
<td>0.06</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>60</td>
<td>3.24</td>
<td>0.05</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>70</td>
<td>3.01</td>
<td>0.08</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>80</td>
<td>2.83</td>
<td>0.17</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>90</td>
<td>2.67</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>100</td>
<td>2.53</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>110</td>
<td>2.43</td>
<td>0.08</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>120</td>
<td>2.32</td>
<td>0.02</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>130</td>
<td>2.24</td>
<td>0.07</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>140</td>
<td>2.16</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>150</td>
<td>2.09</td>
<td>0.04</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>160</td>
<td>2.02</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>170</td>
<td>1.97</td>
<td>0.05</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Baseline [1233]</td>
<td>0.78</td>
<td>0</td>
</tr>
</tbody>
</table>
The MWCV% plot should be considered together with the sex-ratio at length plot. An “optimal” sample size in terms of number of individuals, for which sex should be recorded, can be the number corresponding to a MWCV% sufficiently low and stable e.g. below the 25-th percentiles or below the median value of all the MWCV% values (Figure C.13). It is convenient also to look at the plot of the sex-ratio at length (Figure C.12), with the different hypothesis of sample size in each rectangle.
C4.4 Optimize the number of age readings (script #E)

- Open the the script E_data_simulation_AGE.R and compile the settings indicating the following elements:
  - **myWD**: the path of the working directory, i.e. the path where BioSim Tool folder is located;
  - **caseStudy_path**: the path of the directory of the case study where the input files are located and all the outputs will be saved in the related folders (created by the tool);
  - **species_name**: the name of the species used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - **short_name**: a short code for the analysed species to be used for the output files names;
  - **GSA**: the area used for the selection of the data from the case study dataset (including all the target species and GSAs);
  - **by_cat**: “Y” or “N” indicating if the analyses have to be done considering as sample the trip (if N) or the trip and commercial category (if Y);
  - **y_EUlev6**: “Y” or “N” indicating if the results of the analyses have to be grouped considering the metier/lev6 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - **by_EUlev4**: “Y” or “N” indicating if the results of the analyses have to be grouped considering the gear/lev4 or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - **by_quart**: “Y” or “N” indicating if the results of the analyses have to be grouped considering the quarters or not (notice that this aggregation is used only to summarize the results and not in the analysis);
  - **n_sims**: number of iteration to be run;
  - **samp_sizes**: a vector containing the sample size to be simulated e.g. c(30, 40, …, 200) by length class and by sample;
  - **sampling_design**, a list containing the following elements:
    - **stratified**: TRUE or FALSE indicates if the sampling is stratified (by length class) or not. In script #E it is always TRUE;
    - **strata_var**: the variable for the stratification. In script #E it is always "lenCls";
  - **sampling_options**, a list containing the following elements:
    - **stages**: indicates if the sampling is at one stage or two stages. In the present version of the script only the “one” stage has been tested;
    - **stratified**: TRUE or FALSE indicates if the sampling is stratified or not. In script #E it is always TRUE;
    - **strata_var**: the variable for the stratification. In script #E it is always "lenCls";
    - **stage1_samp_size**: default value NA;
    - **replacement**: indicate if the bootstrap is done with or without replacement. In script #B the bootstrap is always without replacement.
    - **sample_all_available**: default value TRUE;
    - **sample_all_available_warning**: default value TRUE;
    - **vars_to_keep**: filled in only when the sampling is stratified. In script #E it is always c("Age")
  - `load(paste(caseStudy_path, "\input files\input_data.rdata", sep=""))`: the path where the input_data.rdata is located.
# set the working directory
myWD <- "E:\\11_STREAM\\WP3\\TOOLS\\BioSim Tool 1.01\\BioSim Tool 1.01"
setwd(myWD)

caseStudy_path <-"E:\\11_STREAM\\WP3\\CASE STUDIES\\1st_CS7_ANE_PIL_GSA17_18"

species_name <- "Sardina pilchardus"
short_name <- "ANE"
GSA <- "GSA18" # "ALL"

# set sampling design of sample data
sampling_design <- list (stratified = TRUE, strata_var = "lenCls")

# expliciting the number of iterations and the sample sizes (numbers of readings by length class)
n_sims<-10
samp_sizes<-c(seq(2,8, by=1))

# Sampling of different number of individuals without replacement
sampling_options <- list (n_sims = n_sims,     # no of stages
                           stages="one",
                           stratified = TRUE,
                           strata_var = "lenCls", # stratification details
                           stage1_samp_size=NA,
                           samp_sizes = samp_sizes, # samp sizes
                           replacement=FALSE,
                           sample_all_available = TRUE,
                           sample_all_available_warning = TRUE, # replacement options
                           vars_to_keep = c( "Age"))

# read variable table
variable_table <- read.csv2(paste(caseStudy_path,"\\input files\\E_variable_table_biotic_AGE.csv", sep=""), as.is=TRUE)
After the compilation of the settings, run the script #E until the row 162. The summary table with the quantiles of the numbers of individuals by trip/haul will be saved in the output folder.

The summary table with the quantiles can support the selection of the sample size that the trips to be involved in the resampling should have to ensure more reliable results.

Compile the last setting in the script E_data_simulation.R indicating the min_n to be used for the selection of the most representative trips. In the case of simulations in a selected sample size (e.g. 10, 20, ..., 200), please pay attention to select as min_n equal to the maximum of the range (e.g. 200) to allow the script to carry out the resampling without replacement.

After setting min_n at row 162, run again all the script #E to obtain the values of indicators for each sample size.

The outputs tables in .csv format and the plots saved by the script in the folder E_age are below reported.
Table C.36 – Example of the table saved by the script #E. The table reports the annual numbers of individuals for which the age has been recorded in the past years available in the input dataset (from the historical data). In the last row the average calculated over the years is reported.

<table>
<thead>
<tr>
<th>Metier_Gear</th>
<th>quarter</th>
<th>commCat</th>
<th>Year</th>
<th>age_measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014</td>
<td>57</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2015</td>
<td>94</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2016</td>
<td>75</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2017</td>
<td>77</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>2014-2017</td>
<td>76</td>
</tr>
</tbody>
</table>

Table C.37 – Example of the table saved by the script #E containing the values of EMD and the MWCV% calculated as average on the iterations of the total ALK (age-length key) (summed up on all the considered trips). The last line indicates the value of the MWCV% calculated on the original sample (baseline) and the maximum number of the individuals for which the age was read by length class (in brackets). The sample size indicated for the baseline is calculated as average of the number of otoliths by length class in the samples available in the dataset.

<table>
<thead>
<tr>
<th>Metier_Gear</th>
<th>quarter</th>
<th>samp_size</th>
<th>commCat</th>
<th>CV</th>
<th>EMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>5.79</td>
<td>0.06</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>3</td>
<td>-1</td>
<td>4.85</td>
<td>0.04</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>4</td>
<td>-1</td>
<td>4.33</td>
<td>0.02</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>5</td>
<td>-1</td>
<td>4.20</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>6</td>
<td>-1</td>
<td>4.20</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>7</td>
<td>-1</td>
<td>4.20</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>8</td>
<td>-1</td>
<td>4.20</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>Baseline [5]</td>
<td>-1</td>
<td>4.20</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure C.14 – Example of the plots of the age distribution: the proportion (y-axis) of individuals (number of readings by age - on x-axis) calculated applying the ALK. The ALK is calculated as average on the iterations (summing up all the readings by age-length on all the considered trips). In each rectangle a different hypothesis of sample size is done, i.e. a different number of readings is considered by length (2, 3, etc.). The blue lines represent the simulated and the red line the observed in the original sampling.

Figure C.15 - Example of the plot of the MWCV % (y-axis) calculated as average on the iterations of the total ALK (summed up on all the considered trips) considering different sample size, i.e. number of otoliths by length class (x-axis) for each trip/haul. The red line indicates the value of the MWCV% calculated on the original sample (baseline) and the number of the individuals by length class for which the age was read in each trip/haul. The dotted blue lines represent the percentiles (5th, 25th, median, 75th and 95th) estimated on the overall simulations.

The MWCV% plot should be considered together with the ALK plot. An “optimal” sample size in terms of number of individuals for which age should be read can be the number corresponding to a MWCV% sufficiently low and stable e.g. below the 25-th percentiles or below the median of all the MWCV% values (Figure C.15). It is convenient also look at the plot of the age distribution (Figure C.14) with the different hypothesis of sample size in each rectangle.
C5. Case study example

Several applications of the BioSim Tool can be found in the STREAM deliverable 3.3. 
Applications of BioSim were carried out during the WKBIOPTIM3 on the commercial data of M. barbatus in Southern Adriatic Sea, and also in the Aegean Sea (GSA22). The results were compared with the ones obtained with SampleOptimRDBformat script. For details, see Sections 5.3 and 5.6, respectively.

C6. Conclusions

The new version of BioSim, improved and generalized within STREAM project, allows to explore different sampling designs, evaluating the impact of different numbers of individuals collected for maturity, sex and age on the CV and on the corresponding biological estimates (sex ratio, proportion of matures, ALK).
BioSim is based on bootstrap, thus the use of it with limited number of observations could return unreliable results to be considered with caution.

C7. Improvements required

The inclusion of more indicators would undoubtedly be beneficial to provide additional information on the impact of alternative sampling designs on the LFDs.
The development of an R tool including all the scripts of BioSim would be also important to allow a wider availability to the method and to ease the application to other case studies.

C8. GitHub link

https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/BioSim%20Tool%201.01
https://github.com/aleligas/STREAM_MARE-2016-22/tree/Task-3.3-Sampling-Strategy-Optimization

C9. References

D1. Introduction

Development of this tool started under the work package WP3 of the fishPi2 project. WP3 focused on the coordination of the effort of Member States in collecting commercial fisheries data both on-shore all the principal steps involved in the implementation of a sampling design and using a regional dataset and generic simulation code. The aim was to offer a framework that could be used to maximize the efficiency of regional sampling design(s) with respect to pre-defined targets, resources and/or logistical constraints.

The main tool delivered by the fishPi2 project allows the user to test how well a given sampling design can estimate landings, as proxy for other biological parameters of interest. Another branch of this tool was developed to test sampling designs directly on biological parameters, with the example of mean length of species or stocks as the variable of interest.

The user of the tool must define:

- **the population of interest**: for example, all demersal fish in the North Sea, assuming all countries and fleet targeting demersal fish can be sampled
- **the sampling frame**: i.e. all elements that encompass the accessible component of the population and that can be listed (e.g. dates and location of landings, vessels); this information does not exist a priori for fisheries and so, in real applications, last year(s) data can be used assuming they will be roughly similar from one year to the next. The elements of the sampling frame are called PSUs (primary sampling units).
- **the sampling design** = strata and clusters - for example the simulation framework can be used to describe a hierarchical multi-stage stratified random sampling design where the population is divided in non-overlapping groups (strata – one level, non-nested) and, within each stratum, elements of the population are organized in a nested structure (clusters), the definition of this nested structure defining the sampling hierarchy.
  - **strata** = every element of the population (PSU) belongs to a stratum (if the population is stratified) and, in sampling theory, the definition of a stratum should reflect the homogeneity of the metrics of interest, although here the objective might be to logistically implement the sampling design in a cost-efficient manner, which might not be optimal and could introduce bias (e.g. strata based on location within country). Each stratum can have a different target effort. Each sampling unit (PSU) within stratum must however have an equal probability of being sampled in case of a random sampling design. If that is not the case, the implications should be discussed, if not tested.
  - **clusters** = sub-elements of the sampling unit, which are grouped and likely more similar to one another than across the whole population, for example vessels at port, fish within basket. Multi-stage sampling is where there is a hierarchical nesting of clusters (e.g. haul within trip). For example, if PSU is port-day, SSU (secondary sampling unit) can be vessel, TSU (tertiary) can be fish box.
- **the sampling effort**: amount of sampling to be associated with the first stage, i.e. allocation of effort per stratum in terms of number of PSUs to sample, as well as amount of sampling at the 2nd and more levels, i.e. number of SSU per PSU, TSU per SSU. These
choices are to be informed by resources and logistical constraints (e.g. time to process the catch from an entire vessel)

- the domains: these do not influence the sampling design, they are just variables over which one may want to report the output estimates, e.g. combination of species and areas or stocks

In summary, the user must provide a full dataset of the population to sample from (population of interest), a covariate for stratification (if required), the sampling frame or PSU variable, the SSU variable (no TSU option included yet), the effort by stratum (number of PSUs to sample), the effort at PSU level (number of SSUs to sample) and the domains of interest.

The framework assumes a random selection of units at all different stages of sampling. It simulates the sampling one species at a time (i.e. no option for sampling more than one species at once, or concurrent sampling yet). The estimates of population parameters (at population or domains level) are generated using the “survey” package in R, which relies on the design-based Horvitz-Thompson estimator.

The focus of the fishPi development was on mean length but further development for WKBIOPTIM3 made it possible to extract length distributions, with objective to optimise sampling design based on length distribution in future versions (still at an exploratory stage). The tool was built as an R package which is available as a zip file on the GitHub of the project:

https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/FishPi2WP3_BioSims

D2. Libraries required

Relies mostly on the survey package for the survey design-based estimations.

Other packages loaded: ggplot2, grid, Matrix, pracma, dplyr, lubridate, stringdist, stringr, tidyr, reshape2

fishPiCodes, developed under the fishPi project, is also called in (uploaded to GitHub, link provided above).

D3. Input data

The population dataset to sample from is based on the fishPi WP3 data call. A copy of it can be found on the GitHub of the present project. The table below describes the different columns requested, their names and format, with examples.
<table>
<thead>
<tr>
<th>Field</th>
<th>Variable name</th>
<th>Description</th>
<th>Format</th>
<th>Code list or example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sampFishTripId</td>
<td>Unique Fishing trip Identifier.</td>
<td>Character string of length 14.</td>
<td>Examples: SCT20150001, PRT20162474 FRA20156632</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The first 3 characters will be the 3 letter flag country</td>
<td></td>
<td>Not required to match the id in the logbook data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>code, the following 4 will be the year, the last 4 a</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>numeric string with leading zeros.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>sampType</td>
<td>Sampling type</td>
<td>Character string</td>
<td>Code list; at-sea, on-shore</td>
</tr>
<tr>
<td>3</td>
<td>vslFlgCtry</td>
<td>Vessel Flag Country</td>
<td>Character string of length 3.</td>
<td>Fixed code list consisting of: BEL, DEU, DNK, ENG, ESP,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Country code based on ISO 3166 – 1 alpha-3 code.</td>
<td></td>
<td>FRA, IRL, NIR, NLD, PRT, SCT, SWE, WLS, GBI, GBC.</td>
</tr>
<tr>
<td>4</td>
<td>sampDate</td>
<td>Unique sampling date per trip.</td>
<td>YYYY-MM-DD</td>
<td>Finite code list; Examples: 2015-12-05, 2016-02-12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A unique date for the trip - for at-sea the landing or</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>departure date – for on-shore the sampling or landing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>date. Use the same method nationally used when populating</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>the RDB CS trip table.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>loc</td>
<td>Unique location par trip.</td>
<td>Character string of length 5</td>
<td>Fixed code list; Examples: NLIJM, DKTBN, IEKBS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A unique location for the trip - for at-sea the landing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>or departure location – for on-shore the sampling or</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>landing location.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Use the same method nationally used when populating the</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;harbour&quot; field in the RDB CS trip table.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>vslLenCls</td>
<td>Vessel Length Class</td>
<td>Character string of length 6.</td>
<td>Fixed code list consisting of: VL0010, VL1012, VL1218,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vessel length class overall (m)</td>
<td></td>
<td>VL1824, VL2440, VL40XX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DCF LOA classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>area</td>
<td>FAO area codes</td>
<td>Character string</td>
<td>Fixed code list; Examples: 27.4.a, 27.8.c, 27.3.a.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Corresponding to highest possible resolution ICES sub-area</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>area, division.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>foCatEu6</td>
<td>Metier level 6</td>
<td>Character string with gear, target, mesh and selection</td>
<td>Fixed code list; Examples: OTB_DEF_&gt;=120_0_0_GNS_SPF_</td>
</tr>
<tr>
<td></td>
<td></td>
<td>device components, underscore separated.</td>
<td></td>
<td>120-219_0_0</td>
</tr>
<tr>
<td>9</td>
<td>sppCode</td>
<td>Species code - WoRMS Aphia ID</td>
<td>Character string of length 6 or shorter, of numeric</td>
<td>Fixed code list; Examples: 127419, 126436, 11723</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The species codes of all the recorded landings from the</td>
<td>values.</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>sppName</td>
<td>Species name – scientific name. Accepted WoRMS name</td>
<td>Character string</td>
<td>Fixed code list; Examples: Capros aper, Gadus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corresponding to the Aphia ID</td>
<td></td>
<td>morhua, Sepiidae</td>
</tr>
<tr>
<td>Field</td>
<td>Variable name</td>
<td>Description</td>
<td>Format</td>
<td>Code list or example</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>11</td>
<td>catchCat</td>
<td>Catch category. Landing (LAN) or Discard (DIS) or BMS (if any are available)</td>
<td>Character string of length 3.</td>
<td>Fixed code list consisting of: DIS, LAN, BMS</td>
</tr>
<tr>
<td>12</td>
<td>catchWt</td>
<td>The estimated weight of the combined fish in the raised length frequency, whole weight in kg. This is the weight per fishing trip, so will be repeated per length class.</td>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>lenCls</td>
<td>Whole length in mm</td>
<td>Integer</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>estNum</td>
<td>Estimated number of fish per length and fishing trip and the above stratification</td>
<td>Numeric</td>
<td>May be decimal as result of having raised samples to trip level</td>
</tr>
<tr>
<td>15</td>
<td>sampNum</td>
<td>Sum of sampled number of fish per length and the above stratification</td>
<td>Integer</td>
<td></td>
</tr>
</tbody>
</table>

Note that these are actual samples, they were not raised to the population level. The fishPi2 script does not conduct this raising procedure. So simulations from this dataset are based purely on the trips for which length data were collected, i.e. the biological dataset was not linked to the full landings dataset. This would require defining rules on what the samples were thought to represent, which metiers to group etc. This is presumably done on a stock by stock basis prior to including the information in assessments.

**D4 Scripts files**

As stated above, this is an adaptation of the script and functions developed for the main simulations of fishPi2 WP3 to run with biological (lengths) data, improved further under WKBIOPTIM3. It is all in the form of an R package called FishPi4WKBIOPTIM v0.2.0 that can be found on the GitHub, in [https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/FishPi2WP3_BioSims](https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/FishPi2WP3_BioSims)

**D4.1 Functions in the package**

Here is the list of functions found in the package. Each function is documented in the package with input/outputs and examples of how to run them.
D4.2 Wrapper function
This is the main function, which runs the simulations.

```r
Wrapper function to simulate sampling

Description
Wrapper function to simulate sampling

Usage
siml_sampling(dataset =, app, nsim = 20, psu = "fishTripId", myStratum = "random", myEffort = NULL, domains = "area", maxSsuSamp = NULL, replacement = F, concurrent = F, sampleForeign = F, sampCtry = NULL)

Arguments
dataset FishPi2 formatted dataset. See function FishPi2FishPi
app vector of species names
nsim number of simulations
psu options for psu are "fishTripId", "vesselId", "siteXday", with default set at "fishTripId". "fishTripId" is single stage sampling, two-stage sampling with the other options as not been fully tested yet.
myStratum default is "random", or give name of column (as character) that defines strata. For now, either give the name of a variable that exists in the default dataset or create your own new variable and add it as column to the dataset (e.g. combining landing country and area fished and calling it "city_area")
myEffort dataframe with columns as strata and each row is a value of effort to be tested
domains default = "area". Variable for which to summarise the results a posteriori. See svyby from survey package. Note that this variable is not accounted for in the design, i.e. sampling effort is set by strata not by domains and this is not equivalent to post-stratifying as it does not use new weights for the variable
maxSsuSamp max number of SSUs per PSU to sample
replacement sampling with replacement TRUE or FALSE
concurrent NOT IMPLEMENTED YET 6.x. set to TRUE) - concurrent sampling TRUE or FALSE
sampleForeign whether to sample foreign vessels TRUE or FALSE, set to FALSE
sampCtry what country is sampling (if not in the original dataset, provide country 3 letters). This can only accommodate one country, so fix it outside of the function if needs be. The example dataset provided was compiled from submission from different countries and so it already includes sampCtry. If missing, it will use the vessel flag, in which case there will be no option for sampling foreign vessels.
```
**Details**

Read “Value” below first. In the first element of the output list, the summary outputs at the finest level (i.e. mean lengths for a given species at a given sampling effort) contain the following estimates:

- `trialEst` - estimated overall mean length
- `sppPep` - mean length by species in the population
- `RSEst` - sd mean of mean lengths by species (one mean length per simulation)
- `sppRSE` - mean of RSEst
- `meanSppEst` - estimated mean length by species
- `sppEst` - length distributions of each simulation by species
- `sppCIL` - upper confidence interval for mean length by species
- `sppCILU` - lower confidence interval for mean length by species
- `sppSampSize` - sample size by species (number of PSUs) - only for "FishTripod" which always either the psu or ssu
- `sppBiasEst` - bias of the estimate by species, i.e. 100*(meanSppEst/sppPep)-1
- `domPep` - mean length by domain in the population
- `domRSEst` - sd mean of mean lengths by domain (one mean length per simulation)
- `domEst` - length distributions of each simulation by domain
- `domEstsims` - mean length by domain for each simulation
- `meanDomEst` - mean length by domain (averaged over simulations)
- `domCIL` - upper confidence interval for mean length by domain
- `domCILU` - lower confidence interval for mean length by domain
- `domSampSize` - sample size by domain (number of PSUs) - only for "FishTripod" which always either the psu or ssu
- `MWCV_spp` - mean weighted CV of lengths per species, i.e. precision over the entire size range in a length frequency distribution as calculated in BioSim Tool
- `MWCV_domain` - mean weighted CV of lengths per domain, i.e. precision over the entire size range in a length frequency distribution as calculated in BioSim Tool

As the second element of the output list, the frequency data at the finest level (i.e. frequencies for a given species at a given sampling effort) contain matrices of simulated data with length classes as rows and simulation number as columns. Each column sums up to 1 (i.e. frequencies are reported as proportions).

**Value**

Read “Details” above after this. Lists within lists: the first level relates to summary statistics for mean lengths (1st element) or raw length frequencies data (2nd element), the 2nd level splits the outputs by species, the 3rd level splits the outputs by effort level.
D4.3 Example of how to run a simulation script and extract outputs

If starting from RDB data, use the formatRDB2FishPi function to set up the data for use in the package (*function still to be tested*)

```
Description
Format RDB data to run with the FishPi2 simulation package

Usage
formatRDB2FishPi(HL, SL, HH, TR)

Arguments
HL  RDB HL data
SL  RDB SL data
HH  RDB HH data
TR  RDB TR data

Details
Note that these data are not raised to the full population of landings, this is just the sampling data. Starting from RDB data it could be possible to raise the data to the full population using the COST packages. This would require hacking into this function. Then the simulations can be done on either the sampling data or the raised sampling data. FUNCTION NOT YET CHECKED

Value
data formatted in FishPi2 format

Examples

```R
## Not run:
## READ IN RDB data
setwd("TOUR_DIRECTORY")
HL <- read.csv("HL_Eng.csv")
SL <- read.csv("SL_Eng.csv")
HH <- read.csv("HH_Eng.csv")
TR <- read.csv("TR_Eng.csv")

## !!! FUNCTION NOT CHECKED ON DATA OTHER THAN MY OWN !!!
## RUN THE FUNCTION ####
dataset <- formatRDB2FishPi(HL, SL, HH, TR)
head(dataset, 5)
```

## End(Not run)
Here is how to run a full simulation with a range of sampling efforts for a given design and extract outputs for interpretation:

```r
# Not run:
# Select the dataset you want to simulate sampling from
# This dataset should be in FishPak format
# Example dataset containing landing samples of COD, PLE and ODO for meter TBR_DEF_70-59_0_0
data(fishPakFormattedData) # see also help(formatREB2FishPak) function for formatting
select <- fishPakFormattedData

# Run the simulation function
myEffort_nonstrat <- chinl.data.frame(c(5,10,15)) # number of PSUs to sample
out <- simul_sampling(dataset=select, spp=c("COD","PLE"), myEffort= myEffort_nonstrat, domain="area")

# Example of stratified option
myEffort_strat <- chinl.data.frame(q1=seq(2,5,by=2),q2=seq(2,18,by=2),
mystrat = "quarter",
myEffort=myEffort_strat, domain="area", sampleForeign=TRUE, replacement=TRUE)

# Extract specific outputs
# The number in "{i}" refers to the effort level index
out$meanLqthCOD[[1]]$spp$pop
cut(frequencies$COD[[1]])

# Plot RSE
df <- NULL
for (i in linrow(myEffort_nonstrat)){
df <- df %>% chinl.data.frame(spp=="COD", effort=myEffort_nonstrat[i],
                       n=meanLqthCOD[[1]])
}
plot(out$meanLqthCOD[[1]]$spp$pop)

# Plot MCV as a function of sampling effort for COD
df <- NULL
for (i in linrow(myEffort_nonstrat)){
df <- df %>% chinl.data.frame(spp=="COD", effort=myEffort_nonstrat[i],
                       n=meanLqth$COD[[1]]$MCV)
}
plot(out$meanLqth$COD[[1]]$MCV)

# Plot MCV as a function of sampling effort for domains in COD
df <- NULL
for (i in linrow(myEffort_nonstrat)){
  for (k in length(domain)){
    df <- df %>% chinl.data.frame(spp=="COD",
                       n=meanLqth$COD[[1]]$MCV)
  }
}
plot(out$meanLqth$COD[[1]]$MCV)

# Use existing functions to plot results (or make your own)
# Plot based on the mean length outputs
devPlot(out, spp=c("COD","PLE"), idx_effort=1)
devPlot(out, spp=c("COD","PLE"), idx_effort=1)

devPlot(out, spp=c("COD","PLE"), idx_effort=1)

# Plot based on the length frequencites
plot(out$meanLqth$COD[[1]]$spp$pop)

# Plot related to frequencites - STILL WORK IN PROGRESS
plot(out$meanLqth$COD[[1]]$spp$pop)

# End (Not run)
```
The statistical outputs are stored in “out” in the above example.

“out” is a list of outputs of length 2, the first element storing the main statistical outputs based on mean length and the second element storing the length frequencies of all simulations.

These 2 list elements are organized as lists of species, which are themselves organized as lists of sampling effort levels. All the estimates produced, as described in the “Details” section of the wrapper function help file, are therefore produced for each combination of effort level/species in the first element of “out” (i.e. out[[1]] or out$meanLgth), and all simulated lengths distributions are also produced for each combination of effort level/species in the 2nd element of “out” (i.e. out[[2]] or out$frequencies). It is then possible for the user to calculate any other metrics of interest if not already produced in the package.

D4.4 Type of visual outputs
Example of extracting fishPi2 standard outputs and plotting as the user wishes – relative standard error (RSE, i.e. sd/mean) of mean length by sampling effort and species.
Example of extracting outputs and plotting as the user wishes – MWCV of lengths distributions by sampling effort for one species (one value of MWCV for each simulation)

Example of extracting outputs and plotting as the user wishes – MWCV of lengths distributions by sampling effort for one species by domain (one value of MWCV for each simulation)
Example of output of function “devPlot” (deviance plot) for a given sampling effort

Example of output of function “biasPlot” (bias plot) for a given sampling effort
Example of output of function domHist (mean length) for a given sampling effort – in red the mean lengths of the simulations, in green the mean length in the population (original dataset sampled).

Example of output of function freqs_plot (length distribution) for a given sampling effort.
Example of output of function freqs_stats (stats on length distribution to compare simulated lengths data to real data) – temporary function, work in progress.

D6. Conclusions

This package is very much still a work in progress but offers great potential to provide robust statistics out of a flexible framework that allows comparing regional sampling strategies. The code is documented throughout so that any interested party can pick it up and add any layer of complexity they would require to test. The raw format of the outputs also allows the user to make their own visual display or apply their own statistics.

D7. Improvements required

This tool is a work in progress and a number of developments can be made to improve it and broaden its scope:

- check formatRDB2FishPi function to facilitate the formatting of standard datasets
- provide statistical outputs based on length frequency (only exploratory for now)
- make extracting outputs more intuitive
- set option for concurrent sampling (for now it conducts simulations one species at a time)
- make sure it can work with raised data
- run further checks to make sure it does as intended with stratification and with two-stage sampling (all examples worked up with single stage sampling until now)
– add hauls to the level of sampling – at the moment fishing trip is the minimum possible sampling unit and all fish from that trip, available in the dataset, is accounted for (could consider defining number of fish to sample too)
– add a level of sampling to sample age data within length bins (or sample ages directly rather than lengths)
– in general, make the sampling units and their hierarchy more flexible
– add possibility for including finite population correction in the calculation - most relevant when simulating on sample data only as this package currently does make the specification of domains and strata more flexible/generic

D8. GitHub link

https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/FishPi2WP3_BioSims
Annex 4: Slides of oral presentations

1. Nuno Prista (SLU, Sweden): Code developments: optimization at sample-level

Why “optimize”?  
- So that:  
  - We collect only the data we effectively need  
  - We collect more of data that we effectively need  
  - We free ourselves for other work (or free time)  
- End-user oriented “optimization”  
  - Simulate impact of reducing sampling in results at national or international level  
- Data-collector oriented “optimization”  
  - Simulate conservative sampling scenarios that fit different end-users  
  - Easier to grasp, faster to implement, bottom-up approach

Biological Samples  
- Large numbers of biological samples are collected annually in every European fisheries lab (including mine)  
- Biological samples are costly:  
  - The samples themselves  
  - Processing at the lab: measuring, determining maturity  
  - Further processing at the lab: age determination, histology  
  - Logging into national databases and quality checking  
  - But also...  
    - More sampling of one species effectively reduces sampling of other species (contrasts on the sampling of data poor stocks)

Sample-Level vignette  
- 001_prepare_data  
- 002_explore_data  
- 003_sim_data  
- 004_sim_analysis  
- Auxiliary funs and tables  
- Data and result directories
001_prepare_data
- Checks on variables names and formats

002_examine_data
- Inspect frequencies, modes, and sample statistics / indicators
- Select minimum sample size of "representative" samples

003_sim_data
- Simulate different sample sizes (with/without replacement)
- Calculate statistics / indicators for each replicate

004_sim_analysis – "within samples" [a awful name]
- Graph results
  - e.g., CV and MWCV of replicates of each sample and total
- Explore correlations across variables
  - E.g., length and Age
- Select the most appropriate sample size
  - Select a statistic (e.g., CV, MWCV)
  - Select acceptable value for statistic (e.g. 2%)
  - Select a "risk" method (e.g., wos, quantile 0.975, median)
- Determine the sample size that provides statistic at selected "risk"
- Evaluate if impacts on other statistics and variables are acceptable

004_sim_analysis – "modeling" [a bit better name]
- Select a statistic to model
- e.g., CV and MWCV
- Fill exponential models (three types)
- Analyse diagnostics and select a best model
- Predict CV, MWCV across a range of sample sizes
- Multimodel approach: combination of model predictions
  - Normal prediction: Select a sample size -> Select a risk method (e.g., wos, quantile 0.975, median) -> Determine statistic values expected from that sample size with that "risk"
  - Inverse prediction: Select acceptable statistic value -> Select a risk method (e.g., wos, quantile 0.975, median) -> Determine the sample sizes that provide that statistic value with that "risk"
Done
- Streamlined vignettes and functions
  - more general and practical
- Speeded up
  - Implementation in parallel (via snowfall)
- Upload to Github

To be done
- Incorporate recent stats / indicators
- Test with stratified and 2 stage sampling
  - partially developed during whitperm-2
    - Incorporate insights into sample pooling
- Improve graphics and outputs
- Further discuss with users
2. Patricia Gonçalves (IPMA, Portugal): Sampling optimization of Blue whiting - ICES 27.9.a
SamplingOptim R-tool

Suggestion for users:

- Applied it to several years of data;
- Use different scenarios (e.g. year/quarter, sex/ratio variation, Port sampling...)

Optimal number to sample by length class
3. Julia Wischnewski (Thünen Institute of Sea Fisheries, Germany): Quantifying robustness of a length frequency distribution shape.

Main points and motivation
- The overall shape of length frequency distribution (LFD) is the most important attribute (other than descriptive statistics like mean, variance etc.) since it represents the key length patterns displayed by species in an area.
- The problem is wider than estimation of an appropriate sample size, namely: what a robust sampling framework we need for the reasonably precise LFD?
- “Reasonableness” should be understood in both biological and statistical contexts.
- Knowledge of whether annual datasets adequately capture shape variation could improve the practice of collecting length data overall (distribution of the regional sampling effort).

Statistical significance vs biological significance

The definition represents in general one iteration of the iterative subsampling procedure. During the procedure, we remove one length measurement from each length class and check if the requirements above are fulfilled. A subsampling process repeats until they are, otherwise stops. The resampled subsample represents a minimally permitted reference subsample (“the worst case”) that still reveals the patterns of original distributional shape.

Empirical CDFs of original sample and reference subsample are displayed and the L1-distance (i.e. area between CDFs) is calculated. This value we will call the adverse dissimilarity value (ADV).

All subsamples satisfying the requirements above and having an L1-distance between zero and ADV are considered as ensuring a reasonably precise LFD.

Additional optional (L2) restrictions:
- A: a value smoothing the length classes; helps to discover which modality/uni/mode are robust.
- x: a value indicating a minimal required number per length class in subsample.
- P: subsampling permitted length classes.

Example:
1 cm bin width

Example:
3 cm bin width

ADV definition
Let P be an original (parent) sample CDF, G - a subsample CDF, then

\[ ADV = d(P, G) \]

where

\[ d(P, G) = \sum_{i} \min(P_i, G_i) \]

and the requirements are satisfied, \( \forall i \)

Example:
3 cm bin width
Example:
5 cm bin width

Original sample & subsample:
- \( n = 10 \)
- \( n = 5 \)
- All length classes: \( n = 5.6 \)

Robust mode/medians of the model-reduced subsample - original sample

Another example: cod, 2Q 2018, Area 27.4
**Future work direction**

- Revising and improving of algorithm (other dissimilarity measures, bootstrap confidence intervals for original sample CDF preserving correlation structure)
- Regional cooperation in sampling process: recommendations on re-distribution of metiers sampling effort between EU states based on temporal analysis

* Important to note:
  - LFD should be raised to metier level before proceeding!
4. Isabella Bitetto (COISPA, Italy): Sampling optimization and R tools: the contribution of the STREAM project
WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/metros

MAIN OUTCOMES

TASK 3.1
- The data sharing agreement (DSA) finalized in collaboration with the Institutes involved in the STREAM project. The DSA identified the type of data to be shared, the common data format and rules for the access/use of these data.
- A frame detailing and quantifies the type of data needed has been included:
  - sampling data for the optimisation of the sampling intensity in the RSP of 2019 (Task 3.3);
  - age-related data required for the task of agreeing on a set of regional data quality assessments (T 6.2).
- The Data Sharing Agreement has been signed by most NCs (Cyprus, Italy, Slovenia, Greece, Bulgaria, Romania, Spain), white others have ensured the data delivery though they have not signed the DSA (France, Croatia, Malta).
- A data call has been finally launched (with the deadline on July 18).
- Dedicated sections of the STREAM SharePoint have been made available to each NC in order to access and upload the data.

OUTCOMES

TASK 3.2
- Identify solutions for a regional database for MedSSS by defining...

Data type and domain

Sharing and accessibility

Functionality

WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/metros

MAIN OUTCOMES

TASK 3.2
- Develop auxiliary tools (R-scripts) to standardize and ease the procedures for data processing and delivering to specific data calls including the following data transformations.

1. RCG MedSSS Data Cell
2. COST (format SDEF)
3. DG MARE MedSSS Data Cell
4. DG MARE MedSSS Data Cell

WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/metros

OUTCOMES

TASK 3.2
- Update the RCG MedSSS Data Call format for the storage of primary data from commercial fisheries sampling including additional information in CS1 and CS2 tables.

WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/metros

OUTCOMES

TASK 3.2
- Evaluate options regarding the hosting of a RDB against three available scenarios, taking into account the current situation, ongoing studies, developments and suggestions as well as needs from STECF, GFOM and ICES working groups.
  - GFOM - Data Collection Reference Framework (DCRF) for the collection and submission of fishery-related data in the GFOM area (Med 8 DC)
  - providing technical support and resources to start fisheries data collection in many non-EU countries (e.g. Albania, Montenegro, Lebanon, Egypt, etc.) acting in close cooperation with FAO sub-regional projects (ArabMed, MedSudMed, EastMed, CapeMed)...
  - JRC - Data dissemination tool allowing to access the database mainly related to the to the different Data Calls issued by DG MARE:
  - Data are made freely available in aggregated form, after being published in official STECF reports
  - ICES - experience from a new Regional Database and Estimation System (RDBES) currently being developed including statistical estimation methods for the reusing sampling data & new data format providing a common structure to describe both the disaggregated sampling data and, most importantly, how it was sampled (i.e. information on sampling design).
WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/meters

OUTCOMES

TASK 3.3

✓ Perform analyses on sampling optimization in terms of number of individuals to be measured.
✓ Perform analyses on sampling optimization in terms of number of trips.
✓ Perform analyses on sampling optimization on the number for biological measurement.
✓ Perform possible scenarios varying number of trips and measured individuals.

The developed tool consists of two packages: SD Tool v.2 and BiodSim Tool

Statistical principle:
The value of CV decreases with the increase of the number of sampling units, defining a curve. In the SD tool the part of the curve where the tangent changes and begins to flatten (i.e. the curvature range) is considered as a suitable trade-off between the precision and the sampling effort. Then, the sample size (in terms of sampling units corresponding to that part of the curve) is proposed as “optimal” sample size.

The first 4 solutions are considered to find a trade-off among the species.

WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/meters

OUTCOMES

TASK 3.3

e.g. stratification by year (aggregating quarters) by gear level 4 (aggregating meters)

Var1=spatial stratification (e.g. GSA)
Var2=time stratification (e.g. 1-year)
Var3=metric/gear stratification

The BiodSim Tool using bootstrap technique allows to resample the historical data at sample level (i.e. per trip) to identify the minimum no. of fish to be measured in order to avoid oversampling. This minimum number of individuals to be sampled can be used as a threshold for subsampling.
WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/métiers

OUTCOMES

TASK 3.3
- Perform analyses on sampling optimization on the number for biological measures
  - BioSimTool allows to resample by trip:
    - the sex and maturity variables varying the number of individuals measured by trip (e.g. from 20 to 100) as output the sex-ratio at length, the maturity-at-length for each simulation
    - the age variable varying the number of age readings by length class in each trip (e.g. from 2 to 10 by length class) saving as output the age-length keys (ALK) for each simulation

Task 3.3
- Perform possible scenarios varying number of trips and measured individuals
  - Based on the results on the sampling optimization on all the target species (implemented with BioSimTool) and simulations with different number of length measurements (implemented with BioSimTool) the following scenarios can be simulated:
    - Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips) allowing for the same sampling fraction of the stock
    - Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips) collapsing age class means of the age distribution
    - Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips) collapsing age class means of age distribution
    - Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips) collapsing age class means of age distribution
    - Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips) collapsing age class means of age distribution

<table>
<thead>
<tr>
<th>Type of scenario</th>
<th>Description of scenario</th>
<th>Change in number of trips</th>
<th>Change in number of measured individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Simulate different number of trips taken from the <code>historical</code> sampling range (e.g. from 5 to 10 trips)</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>Simulate different number of trips taken from the <code>historical</code> sampling range (e.g. from 5 to 10 trips)</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>Simulate different number of trips taken from the <code>historical</code> sampling range (e.g. from 5 to 10 trips)</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>Simulate different number of trips taken from the <code>historical</code> sampling range (e.g. from 5 to 10 trips)</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

Notes:
- Y = Yes
- N = No
- Scenario 1: Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips)
- Scenario 2: Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips)
- Scenario 3: Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips)
- Scenario 4: Simulate different number of trips taken from the `historical` sampling range (e.g. from 5 to 10 trips)
WP3 - A regional sampling plan (RSP) for 2019 covering commercial fisheries/stocks/mixtures

OUTCOMES

TASK 3.3

The main output of a simulated scenario for each species consists of:
- Total CV
- Mean length distribution for the whole population

To express the performance of each scenario, the values obtained for each species of two indicators will be compared to the respective values of each species in the baseline:
- Coefficient of Variation (CV)
- Earth Mover’s Distance (EMD)

A summary of changes compared to the past is done in order to have a complete overview by species and scenario.

1st Training Workshop on sampling optimization

Venue: 18-21 September 2018, Kavala (Greece)

Participants: 13 scientists from Cyprus, France, Greece, Italy, Romania and Spain

Case studies started during the workshop:

<table>
<thead>
<tr>
<th>Case study</th>
<th>Area</th>
<th>Grids</th>
<th>Countries</th>
<th>Fisheries</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>North Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>2</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>3</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>4</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>5</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>6</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
</tbody>
</table>

2nd Training Workshop on sampling optimization

Venue: 1-4 October 2018, Bari (Italy)

Participants: 9 scientists from 4 countries – Bulgaria, Croatia, Italy and Spain

Case studies started during the workshop:

<table>
<thead>
<tr>
<th>Case study</th>
<th>Area</th>
<th>Grids</th>
<th>Countries</th>
<th>Fisheries</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>North Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>2</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>3</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>4</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>5</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
<tr>
<td>6</td>
<td>Black Sea</td>
<td>20</td>
<td>Bulgaria, Romania</td>
<td>B30 (0.45)</td>
<td>A. diana (7)</td>
</tr>
</tbody>
</table>

1st and 2nd Training Workshop on sampling optimization

Work done:
- Participants worked in groups for training on the use of the scripts.
- A case study was assigned to each group.
- Preliminary investigation on data (data check, coverage, robustness, etc.), and preliminary runs for sampling optimization in terms of number of trips and in terms of individuals to be measured have been done.
- The work undertaken by participants during the workshop will continue “offline” by emails and skype meetings.

Dataset: data have been checked and it seems that a quite good coverage is available for the case study purposes.

An overview on a subset of the case studies, started during the training workshops and subsequently completed, follows:

E. encrasulus and S. pilchardus in GSA 1-5-6-7

Encarmen Garcia and Miguel Vivas (IEO)
Case study 1
E. encrasicolus and S. pilchardus in GSAs 1-5-6-7

Simulations:
- Baseline (average available years)
- Scenario 1: # trips as baseline and with sub-sample (1/2), applied only in cases of representative samples
- Scenario 2: # trips = minimum optimized and no sub-sample
- Scenario 3: # trips = minimum optimized and with sub-sample (1/2), applied only in cases of representative samples
- Scenario 4: # trips doubled respect to the minimum in the optimized range and with sub-sample (1/2), applied only in cases of representative samples

For all the combinations species-GSA the alternative simulated scenario returns a precision comparable to the baseline, except for E. encrasicolus in GSA 9 and S. pilchardus for GSA 7 (the best was in both cases Scenario 4).

This would imply a number of trip 3 times and a consequence of measurements increased respect to the baseline for GSA 5 and an increase of 20% for GSA 7.

Case study 2
A. foliacea and A. antennatus in GSAs 9-10-11

Simulations:
- Baseline (average available years)
- Scenario 1: # trips as baseline and with sub-sample (1/2), applied only in cases of representative samples
- Scenario 2: # trips = minimum optimized and no sub-sample
- Scenario 3: # trips = minimum optimized and with sub-sample (1/2), applied only in cases of representative samples
- Scenario 4: # trips doubled respect to the minimum in the optimized range and with sub-sample (1/2), applied to all the samples

For GSA 9 for A. foliacea the Scenario 4 was the best performing for OTB_MOD, while it was Scenario 3 for OTB_DWS. For A. antennatus it was the baseline for OTB_MOD and Scenario 4 for OTB_DWS.

For GSA 10 and GSA 11 for all the meter and both species, Scenario 4 was the best performing.

For GSA 10 Scenario 4 foreseen about double number of trip by meter respect to baseline, while for GSA 11 an increase of 83%. For GSA 9 Scenario 3 and 4 are associated to an increase of 76% of trip for OTB_MOD and of 100% for OTB_DWS.

Case study 3
E. encrasicolus and S. pilchardus in GSAs 17-18

Simulations:
- Baseline (average available years)
- Scenario B1: # trips = minimum optimized by quarter and no sub-sample
- Scenario B2: # trips = minimum optimized by quarter and sub-sample
- Scenario B3: # trips = minimum optimized by year, equally distributed in the quarters and no sub-sample
- Scenario B4: # trips = minimum optimized by year, equally distributed in the quarters and sub-sample
- Scenarios C1, C2, C3, C4: as B1, B2, B3 and B4 with sub-sample

Although other scenarios show better results in terms of sample either E. encrasicolus or S. pilchardus, they lead to an optimal compromise between the two species for Croatia.

For the Italian PTM fleet (GSA7 and GSA18), there is an increase in the number of trips according to scenario B1 would have a very small effect on S. pilchardus, with a decrease in the CV of less than 10%.

Case study 4
M. barbatus in GSA 29

Cosmidano Neglia (COISPA), Magda Nenciu George Tiganov (NIMRD)

Data available:
- 2016
- Bulgaria, Romania

Stratification of optimization and simulations:
- # trips
- # length measurements
- # sex and maturity measurements

For the Italian PTM fleet (GSA7 and GSA18), there is an increase in the number of trips according to scenario B1 would have a very small effect on S. pilchardus, with a decrease in the CV of less than 10%.
Case study 4
M. barbatus in GSA 29

Simulations:
- Scenario 1: 6 trips = maximum of the optimized range and with subsampling (x1/2).
- Scenario 2: 6 trips = baseline, with subsampling (x1/2).
- Scenario 3: 6 trips = maximum of the optimized range, and with subsampling (x1/2), applied to all samples.

Scenario 3, suggesting a significant increase (40%), in the number of length measurements (about 50%), is the best performing.

---

Case study 5
M. merluccius, M. barbatus and A. foliacea in GSAs 22-23 and 25

Data available:
- Cypriot catches and data (HCMR) and Stratis Katsanos (DTAF)

Greece

Stratification of optimization and simulations:

SD Tool

No improvement on precision would be obtained applying an increase of the number of trips in GSA 25 and a decrease in GSA 22, as suggested by the initial results.

For GSA 25, following the results from the optimization analysis, it is proposed to increase the sampling intensity in terms of sampled trips for GTB and MF65 fisheries, as in Scenario 3. An increase is also proposed on the number of sampled individuals by trip for M. barbatus, for both fisheries.

---

BioSim Tool

Conclusions

The features of the tool developed by STREAM project that are considered important in the perspective to create a unique tool for sampling optimization are:

- The possibility to simulate the impact on precision of several sampling designs (simple random, stratification by meter, quartet, GSA, country or combinations);
- The impact on precision is evaluated on the raised distributions, actually the ones used in the stock assessment;
- Simulations can be carried out in the hypothesis of subsampling, that can be differentiated by stratum (e.g. meter) and also by commercial category.
5. Isabella Bitetto (COISPA, Italy): Quality indicators: the contribution of the STREAM project
Detection of outliers

The report informs about possible lengths outside the allowed range, indicating the trip code where were detected the outliers.

Consistency of age data

The code foresees a visual check of the age-length data by sex and by year combined with the indication of the trip where is the outlier.

A summary table is also produced, reporting the number of individuals sampled for the age by length class and year.

Consistency of maturity data

Visualization of the length distribution by maturity stage and year.

A summary table to evaluate the number of individuals sampled for the maturity by length class and year is also produced.

Consistency maturity data respect to size at first capture

The code carries out the estimation of the maturity ogive curve by sex, allowing the comparison of the maturity parameters with the ones estimated in other areas and in bibliography.

A priori QC on sample biological data in RCG_CS

An R script aimed at producing an automatic report file, containing the outcomes of several essential quality checks on the sampling data (R markdown).

A number of quality checks were foreseen to ease the identification of the most common errors in the sampling data, before the raising procedures.

- Consistency of LFD by year;
- Consistency with allowed ranges of length;
- Consistency of LFD by year and commercial category;
- Consistency of length-weight relationship;
- Consistency with allowed ranges of weight;
- Consistency of age-length relationship;
- Consistency with allowed range of age;
- Consistency of sex and maturity stages.

Consistency of LFD by year and commercial category

Visual checks to easily detect outliers.
A posteriori QC on DGMARE_MedBS tables

Using the conversion tools developed within WP1 (task 3.2), the sampling data can be converted from the RCG_CS in COST and then, in DGMARE_MedBS datasets format.

These formats were chosen for the "a posteriori" checks because it can be converted in the formats required by FDI and GPMC datasets (see the other tools developed in task 3.2).

The checks implemented on these tables concern the temporal and spatial coverage of the following tables:
- Catch at age (landing and discard);
- Landing by length;
- Discard by length;
- Maturity at length;
- Maturity at age;
- Sex ratio at length;
- Sex ratio at age;
- ALK;
- Growth parameters.

Spatial and temporal coverage

Quality checks on DGMARE_MedBS data file formats

<table>
<thead>
<tr>
<th>Coverage</th>
<th>The coverage can be evaluated also by meter, country and species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catch</td>
<td>Checking the file &quot;Catch_2015.xls&quot;</td>
</tr>
<tr>
<td>Table 1) landing at age coverage</td>
<td></td>
</tr>
<tr>
<td>COUNTRY</td>
<td>YEAR</td>
</tr>
<tr>
<td>ITA 2014</td>
<td>1</td>
</tr>
<tr>
<td>ITA 2016</td>
<td>1</td>
</tr>
<tr>
<td>ITA 2017</td>
<td>1</td>
</tr>
<tr>
<td>ITA 2014</td>
<td>1</td>
</tr>
<tr>
<td>ITA 2014</td>
<td>1</td>
</tr>
<tr>
<td>ITA 2015</td>
<td>1</td>
</tr>
<tr>
<td>ITA 2016</td>
<td>1</td>
</tr>
<tr>
<td>ITA 2017</td>
<td>1</td>
</tr>
</tbody>
</table>

Spatial and temporal coverage of biological parameters

On biological parameter tables the coverage checks foresee visual and tabular outcomes.

Sum of products

Specifically for catch at age table, the check concerning the consistency between the sum of products (nb of the individuals raised to the power of the landing/discard" individual weight) and the landing/discard.

Link with FishPi²

After a meeting in Port de Beràsia (October 2015) between WP6 STREAM and FishPi² scientists, an attempt was made to apply the same functions of the CLEFRD library, recently developed by FishPi², to the RCG CS format to create simple Time, Landing and Space objects and allow how to apply that library also in Mediterranean context.

Conclusions

This tool developed in task 6.1 implements a set of essential quality checks on the sampling data and on the DGMARE_MedBS tables that could represent a procedure common and harmonized among the MS and between Mediterranean and Black Sea.

The use of R as programming language makes quite straightforward the inclusion of other checks in a shared quality check procedure.

The possibility of quickly get an automatic report on standard formats, facilitates the use of these tools in the session of data preparation generally preceding the actual stock assessment session in the relevant Mediterranean and Black Sea working groups.
Annex 5: Scripts and Code

The scripts used during the workshop for the case studies are available in Github:

SampleLevelOptim:
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/sampleleveloptim

BioSimTool:
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/BioSim%20Tool%201.01
https://github.com/aleligas/STREAM_MARE-2016-22/tree/Task-3.3-Sampling-Strategy-Optimization

SampleOptim:
https://github.com/gonpatricia/SampleOptimRDBformat
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SampleOptim

SampleReferenceLevel:
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/ADV

Multi-level analysis of sampling effort needed to produce estimates of catch numbers for stock assessment (SimPop):
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/simpop

SDTOol:
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/SD%20Tool%202.04

FishPi4WKBIOPTIM:
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/FishPi2WP3_BioSims

Investigate ALK effect on age distribution in a virtual environment (sampalk)
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/sampalk

LanAge_opt (DRAFT):
https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/lanAge_opt
Annex 6: Draft proposal for standardized notation

Note: this draft proposal was elaborated and presented in the aftermath of WKBIOPTIM3 meeting and should be considered provisional until it is fully discussed by WKBIOPTIM4.
set.seed(1)
box.mnorm(100, 5, 1)

Generic

$S_s$ Sample s (s=1,2,...,S) from infinite population

$SubS_{S, SRSWR, n}$ SubSample from s using SRSWR with size n (where n in 1..N)

$SubS_{S, SRSWR, n,r}$ SubSample from s using SRSWR with size n replicate r (where
r=1..R)

$SubS_{S, SRSWR, n,r,b}$ SubSample from s using SRSWR with size n replicate r,
bootstrap b (b=1..B)

Example

$S_1$ Sample s=2 from infinite population

$SubS_{1, SRSWR, 10}$ SubSample from s=1 using SRSWR with size n=10

$SubS_{1, SRSWR, 10, 3}$ SubSample from s=1 using SRSWR with size n=10 replicate r=3

$SubS_{1, SRSWR, 10, 3, (S)}$ SubSample from s=1 using SRSWR with size n=10 replicate
r=3, bootstrap b
Simulations: Generic

**SRSWR**

\[
\begin{align*}
\text{SubS}_1 \text{SRSWR}_1, n_1 \\
\text{SubS}_2 \text{SRSWR}_2, n_2 \\
\vdots \\
\text{SubS}_U \text{SRSWR}_U, n_U \\
\end{align*}
\]

**SRSWOR**

\[
\begin{align*}
\text{SubS}_1 \text{SRSWOR}_1, n_1 \\
\text{SubS}_2 \text{SRSWOR}_2, n_2 \\
\vdots \\
\text{SubS}_U \text{SRSWOR}_U, n_U \\
\end{align*}
\]

Where: \( \text{SubS}_{\text{SRSWR or SRSWOR}} = (x_1, x_2, \ldots, x_n) \) selected by SRSWR or SRSWOR.

---

**SRSWR**

\[
\begin{align*}
\bar{x}_{\text{SRSWR}} &= \frac{\sum x_{\text{SRSWR}}}{n} \\
\text{CV}(x)_{\text{SRSWR}} &= \frac{\text{se}(x)_{\text{SRSWR}}}{\bar{x}_{\text{SRSWR}}} \times 100 \\
\end{align*}
\]

Where:

- \( \bar{x}_{\text{SRSWR}} = \frac{\sum x_{\text{SRSWR}}}{n} \) Mean of 1st replicate of SubSample of size \( n \) from \( s \)
- \( \text{CV}(x)_{\text{SRSWR}} = \frac{\text{se}(x)_{\text{SRSWR}}}{\bar{x}_{\text{SRSWR}}} \times 100 \) CV of the mean of 1st replicate of SubSample of size \( n \) from \( s \)
- \( \text{se}(x)_{\text{SRSWR}} = \sqrt{\frac{s^2(x)_{\text{SRSWR}}}{n}} \) Standard error of the mean of 1st replicate of SubSample of size \( n \) from \( s \)
- \( s^2(x)_{\text{SRSWR}} = \frac{\sum (x_{\text{SRSWR}} - \bar{x}_{\text{SRSWR}})^2}{n-1} \) Variance of 1st replicate of SubSample of size \( n \) from \( s \)
- \( x_{\text{SRSWR}} \) \( x_{\text{SRSWR}} \) x_{\text{SRSWR}}

---

**SRSWOR**

\[
\begin{align*}
\bar{x}_{\text{SRSWOR}} &= \frac{\sum x_{\text{SRSWOR}}}{n} \\
\text{CV}(x)_{\text{SRSWOR}} &= \frac{\text{se}(x)_{\text{SRSWOR}}}{\bar{x}_{\text{SRSWOR}}} \times 100 \\
\end{align*}
\]

Where:

- \( \bar{x}_{\text{SRSWOR}} = \frac{\sum x_{\text{SRSWOR}}}{n} \) Mean of 1st replicate of SubSample of size \( n \) from \( s \)
- \( \text{CV}(x)_{\text{SRSWOR}} = \frac{\text{se}(x)_{\text{SRSWOR}}}{\bar{x}_{\text{SRSWOR}}} \times 100 \) CV of the mean of 1st replicate of SubSample of size \( n \) from \( s \)
- \( \text{se}(x)_{\text{SRSWOR}} = \sqrt{\frac{s^2(x)_{\text{SRSWOR}}}{n}} \) Standard error of the mean of 1st replicate of SubSample of size \( n \) from \( s \)
- \( s^2(x)_{\text{SRSWOR}} = \frac{\sum (x_{\text{SRSWOR}} - \bar{x}_{\text{SRSWOR}})^2}{n-1} \) Variance of 1st replicate of SubSample of size \( n \) from \( s \)
- \( x_{\text{SRSWOR}} \) x_{\text{SRSWOR}} x_{\text{SRSWOR}}
Simulations: Generic (with boot step)

SRSWR

Boot step

Replicate results

Where:

\[ x_{\text{SRSWR}} = \frac{\sum_{i=1}^{n} x_{i}}{B} \]

\[ s^2(x)_{\text{SRSWR}} = \frac{\sum_{i=1}^{B} (x_i - x_{\text{SRSWR}})^2}{B - 1} \]

\[ se(x)_{\text{SRSWR}} = \sqrt{s^2(x)_{\text{SRSWR}}} \]

\[ CV(x)_{\text{SRSWR}} = \frac{se(x)_{\text{SRSWR}}}{x_{\text{SRSWR}}} \times 100 \]

Mean of 1st bootstrap of replicate i-1 from SubSample of size n from s

Variance of the mean of the 1st replicate of SubSample of size n from s

Standard error of the mean of 1st replicate of SubSample of size n from s

CV of the mean of 1st replicate of SubSample of size n from s

SRSWR

Boot step

Replicate results

Where:

\[ x_{\text{SRSWR}} = \frac{\sum_{i=1}^{n} x_{i}}{B} \]

\[ s^2(x)_{\text{SRSWR}} = \frac{\sum_{i=1}^{B} (x_i - x_{\text{SRSWR}})^2}{B - 1} \]

\[ se(x)_{\text{SRSWR}} = \sqrt{s^2(x)_{\text{SRSWR}}} \]

\[ CV(x)_{\text{SRSWR}} = \frac{se(x)_{\text{SRSWR}}}{x_{\text{SRSWR}}} \times 100 \]

Mean of 1st bootstrap of replicate i-1 from SubSample of size n from s

Variance of the mean of the 1st replicate of SubSample of size n from s

Standard error of the mean of 1st replicate of SubSample of size n from s

CV of the mean of 1st replicate of SubSample of size n from s
Annex 7: LANEAG_E_OPT

Draft tool based on WKBIOPTIM3 (to be finalised in the next meeting)

Eirini Mantzouni (FRI, Kavala, Greece)
emantzo@inale.gr

1. Introduction

This tool is largely based on the “sampalk” script developed during WKBIOPTIM3: “Sampling strategies for age composition in commercial sampling and its effects on age distribution estimates” (see https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/sampalk). 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 finalised 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 LANEAG_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 analyzed 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 dist
library(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 Section 3.1 above.

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 with 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 (e.g., sex, quarter)
- Figure showing landings age distribution by gear and other levels of stratification specified (e.g., 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 (e.g., 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 Section 5.6.

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 (e.g., 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 next meeting. The main points for the future work will include:

- presentation, discussion and testing of the tool (here presented in a draft version) regarding potential improvements and quality indicators used
- potential adoption of the SampleReferenceLevel approach
- apply to additional case studies and scenarios
8. GitHub link

https://github.com/ices-eg/wk_WKBIOPTIM3/tree/master/lanAge_opt