FOURTH WORKSHOP ON OPTIMIZATION OF BIOLOGICAL SAMPLING (WKBIOPTIM4; OUTPUTS FROM 2021 MEETING)

VOLUME 4 | ISSUE 69

ICES SCIENTIFIC REPORTS

RAPPORTS
SCIENTIFIQUES DU CIEM



ICES INTERNATIONAL COUNCIL FOR THE EXPLORATION OF THE SEA
CIEM CONSEIL INTERNATIONAL POUR L'EXPLORATION DE LA MER

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H.C. Andersens Boulevard 44-46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

ISSN number: 2618-1371

This document has been produced under the auspices of an ICES Expert Group or Committee. The contents therein do not necessarily represent the view of the Council.

© 2022 International Council for the Exploration of the Sea

This work is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). For citation of datasets or conditions for use of data to be included in other databases, please refer to ICES data policy.



ICES Scientific Reports

Volume 4 | Issue 69

FOURTH WORKSHOP ON OPTIMIZATION OF BIOLOGICAL SAMPLING (WKBIOPTIM4; OUTPUTS FROM 2021 MEETING)

Recommended format for purpose of citation:

ICES. 2022. Fourth Workshop on Optimization of Biological Sampling (WKBIOPTIM4; Outputs from 2021 meeting).
ICES Scientific Reports. 4:69. 80 pp. http://doi.org/10.17895/ices.pub.21103000

Editors

Isabella Bitetto • Gwladys Lambert • Patrícia Gonçalves

Authors

Isabella Bitetto • Mollie Elizabeth Brooks • Nicholas Carey • Jessica Craig • Ana Cláudia Fernandes • Patrícia Gonçalves • Kirsten Birch Håkansson • Gwladys Lambert • Ioannis Ntokos • Danai Mantopoulou Palouka • Julia Wischnewski

ICES International Council for the Exploration of the Sea
CIEM Conseil International pour l'Exploration de la Mer

Contents

i Executive summary ... ii
ii Expert group information ... iii
1 Introduction ... 1
1.1 WKBIOPTIM 4 participants and agenda ... 1
1.2 Background to WKBIOPTIM 4 and work organization ... 1
1.3 R-tools description and new developments ... 2
1.3.1 Admissible Dissimilarity Value (ADV) (JW) ... 2
1.3.2 STREAMline (IB) ... 5
1.3.3 SampleOptim (PG) ... 6
1.3.4 FishPi2 (GL) ... 8
1.4 Quality indicators ... 8
1.5 R-package ... 11
1.5.1 Process ... 11
1.5.2 Functions description ... 11
2 RDBES ... 14
2.1 Brief overview of RDBES data model ... 14
2.2 Transition to RDBES model ... 21
2.3 Linkage of the different hierarchies to the R-tools ... 21
2.3.1 Stratification ... 25
3 Sampling with and without replacement ... 26
3.1 Literature review ... 26
3.2 Case studies applying a subsampling with and without replacement ... 27
4 Case studies ... 28
4.1 Admissible Dissimilarity Value (ADV) – The case study of Red Mullet in the Aegean Sea (GSA 22) ... 28
4.1.1 Construction of the reference subsample ... 28
4.1.2 Reducing sampling effort by eliminating sampling units (fishing trips) ... 32
4.2 SampleOptim – The case study of Blue Whiting in ICES Division 27.6.a, a comparison of sampling with or without replacement ... 32
4.2.1 Methods ... 33
4.2.2 Results and Discussion ... 35
4.3 Simulated population - applied to two different subsampling designs (with and without replacement) ... 41
4.3.1 Methods ... 41
4.3.2 Results ... 42
4.3.3 Discussion ... 43
4.4 BioSim Tool – The case study of *Mullus barbatus* in GSA 22 ... 44
4.4.1 Methods ... 44
4.4.2 Results and discussion ... 44
5 Next steps ... 48
6 References ... 49
Annex 1: List of participants ... 50
Annex 2: Agenda ... 51
Annex 3: Presentations ... 53
Annex 4: R-function to transform RDBES data tables to RDB format ... 76

i Executive summary

The workshop on optimization of biological sampling (WKBIOPTIM4) was the fourth meeting of a series of workshops aiming at collaborating on fish sampling optimisation processes across ICES Member States. It aimed primarily at providing an update on the development of the different simulation approaches presented and tested during the third workshop (WKBIOPTIM3), at working on shared indicators across these tools, and at creating an R package for end users.

Multiple simulation approaches were discussed during the first three workshops, developed as R scripts, and also contributed by separate projects (STREAM and FishPi²). The focus of WKPBIOPTIM4 was on the following tools: [1] two STREAM tools, BioSimTool and SDTool; [2] the Fishpi4WKBIOPTIM package; [3] SampleReferenceLevel (ADV); and [4] SampleOptim. While there is no current follow-up on the FishPi² project that the group was made aware of, progress has been made on the other approaches aforementioned: a project called STREAMline is expected to build upon the outputs of STREAM; SampleOptim is being further developed and applied in Portugal; and the SampleReferenceLevel (ADV) approach was published (Wischnewski *et al.*, 2020).

Over the week of the workshop, one subgroup focused on the development of an R-package for indicators, with the objective of making the outputs of simulations comparable across approaches and facilitating interpretation for end-users through documentation. A second subgroup worked on testing the comparativeness of these tools in order to feed into this process of comparison and interpretation across tools. Additionally, the effect of simulating sampling with and without replacement on model outputs was discussed and some investigation was conducted in two case studies: one using the SampleOptim tool with blue whiting data and the other using data from a simulated population. Finally, a third subgroup focused on a significant upcoming change which will affect the data input of the tools being developed, i.e. the move from the current Regional Database (RDB) to the new Regional Database and Estimation System (RDBES).

Future work on the WKBIOPTIM tools should include continuing development and testing, the R-package development, and adaptations to accommodate the sampling schemes from the different RDBES hierarchies.

ii Expert group information

Expert group name	Workshop on Optimization of Biological Sampling (WKBIOPTIM 4)
Expert group cycle	Annual
Year cycle started	2021
Reporting year in cycle	1/1
Chairs	Isabella Bitetto, Italy
	Gwladys Lambert, UK
	Patrícia Gonçalves, Portugal
Meeting venue and dates	15-19 November 2021, online (11 participants)

1 Introduction

The Fourth Workshop on Optimization of Biological Sampling (WKBIOPTIM 4) chaired by Isabella Bitetto (Italy), Gwladys Lambert (UK) and Patrícia Gonçalves (Portugal) met *online*, from 15th to 19th of November 2021, to:

a) Develop further indicators of length and age frequency data by i) testing the different indicators and quality thresholds using simulations and ii) preparing an R package with the functions used to calculate them; (Science Plan codes: 3.3);
b) Consolidate and update existing open source code used in previous workshops (BIOPTIM1-3) and generalize for wider use, package code and document tools, and assess compatibility of tools with use of standard data formats and sources; (Science Plan codes: 3.2);
c) Continue to provide support on the use of WKBIOPTIM tools with the aim of a future optimization at national/stock/regional levels. (Science Plan codes: 3.2 and 3.3).

1.1 WKBIOPTIM 4 participants and agenda

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

1.2 Background to WKBIOPTIM 4 and work organization

The WKBIOPTIM series aims to develop tools in R 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 the lower stages of national sampling programs (e.g., number of trips, hauls within trips, fish within hauls), combined with inefficient sampling effort distribution, may not be providing significant additional information on the population. In response to this, some national labs started developing statistical tools with the aim of analysing and optimizing biological sample sizes. The aim was to assess the impact of reducing sampling effort on age or length distribution estimates in clear-cut cases of excessive sampling, or of increasing sampling effort where the information collected may not be sufficient. This work was presented at the first WKBIOPTIM and jointly developed by its participants thereafter.

All WKBIOPTIM R scripts are based on commercial sampling data extracted in the RDB format[1]. Scripts for conversion of the DATRAS format to the RDB format have also been developed to allow some of the WKBIOPTIM tools to run on research survey data (ICES, 2019a). Overall, the main part of the preparation/development of the scripts was 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 during/after the WK. The code is already being used by some institutes to assess potential for improvements in sampling (e.g. by reducing oversampled species or evaluate how to increase sampling in other that are under sampled) (ICES, 2017, ICES, 2019a, ICES, 2019b). The R scripts were made available to the participants through the GitHub (ICES, 2019b) or the Sharepoint. The work performed during the workshops has been presented in several groups and ICES working groups (STECF, WGBIOP,

[1] https://www.ices.dk/marine-data/Documents/RDB/RDB%20Exchange%20Format.pdf

WGCATCH, etc.) and have received positive feedback and good incentives to continue. However, in order for the WKBIOPTIM tools to be useful in the long-run, the data inputs will have to move from the RDB format to RDBES. This will further add the possibility to make use of the information contained in RDBES regarding the sampling designs in the simulations.

The fourth WKBIOPTIM aimed to continue working on the R scripts presented in previous workshops. During the workshop, the practical work was divided between three subgroups:

1. Developers: this group started to format the tools for compatibility within an R package, together with documenting the functions; also started to develop functions for quality indicators [TOR a); TOR b)].
2. Testing group: this group tested the tools available in WKBIOPTIM3 on other case studies and gave ideas on possible integrations and improvements. Sampling with and without replacement was also tested [TOR a); TOR b)].
3. RDBES group: this group focused on the steps towards using the RDBES format [TOR c)].

1.3 R-tools description and new developments

Updates on new developments of a number of R tools considered in previous workshops were presented to the group. Those were: Admissible Dissimilarity Value (ADV or SampleReferenceLevel) (Section 1.4.1); STREAMline optimization tools (Section 1.4.2); SampleOptim (Section 1.4.3); and FishPi[2] (Section 1.4.4).

The slides from the presentations are included in Annex 3.

1.3.1 Admissible Dissimilarity Value (ADV)[2] (JW)

The proposed framework (ADV or SampleReferenceLevel), as well as the corresponding R tool, was discussed at the third Workshop on Optimization of Biological Sampling (WKBIOPTIM3) (ICES, 2019b), and then published in Wischnewski et al. (2020). The approach aims to identify a reduced but still informative sample (subsample) and to quantify the (dis)similarity between reduced and original samples. At the core of the approach is the concept of reference, or benchmark, subsample, which is the minimal representative subsample preserving a reasonably precise length frequency distribution (LFD) for a selected species. An iterative deterministic subsampling procedure, based on defined conditions, returns a reference subsample, quantifies the difference between the original sample and the reference subsample and provides a threshold value. This threshold is called an admissible dissimilarity value (ADV).

The LFD always displays a range of modal length classes or modes (bumps, spikes) and antimodal length classes or antimodes (gaps, dips). Generally, LFD is difficult to quantify. The standard bandwidth Δ = 1 cm, recognized in standard Regional Database (RDB) Format, obviously delivers the maximal number of modes and antimodes present in the data set, and can cause some "spurious" modes and antimodes. The definition below helps to find a formal way to verify the dissimilarities between LFDs of original sample and subsample.

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

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

[2] SampleReferenceLevel

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

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

(1) $A_r^{\Delta} \leq A_j < A_r^{\Delta} + \Delta$
(2) $A_j = \min_{(A_1, A_2, \dots) \in [A_r^{\Delta}; A_r^{\Delta} + \Delta[} (A_1, A_2, \dots)$.

This definition helps to identify the so-called robust modes and antimodes, which continue to be present in the subsample despite length class smoothing and, therefore, are not suspect to sampling artefacts, "contaminating" the distributional shape.

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

Definition 2. Let $\vec{M} = (M_1, M_2, \dots)^T$ and $\vec{m} = (m_1, m_2, \dots)^T$ be robust modes and $\vec{A} = (A_1, A_2, \dots)^T$ and $\vec{a} = (a, a_2, \dots)^T$ be robust antimodes of LFD of the original $S_{orig} = S_0$ and reduced S_n samples, respectively. We define S_0 and S_n as similar, if:

(1) they have the same number of robust modes and antimodes revealed under chosen bandwidth Δ, 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 $\varepsilon = \varepsilon$ (max species length)
(3) amplitudes ratio $\frac{|g(m_i) - g(a_j)|}{|f(M_i) - f(A_j)|} \geq \theta$, where $f(\cdot), g(\cdot)$ are the values of the original and reduced sampled LFDs at a point, respectively; $j \in \{i;\ i+1\}$, $i \in \mathbb{N}$, $0 < \theta \leq 1$.

Roughly speaking, this definition states that the subsampled data set has to preserve the structure and specific patterns of the original data set, namely: 1) reveals the same number of robust modes and antimodes; 2) allows the locations of modes and antimodes for larger specimens to vary in some small interval defined by parameter ε; 3) keeps distinguished differences between adjacent modal/antimodal values, controlled by the parameter θ. If conditions (1)-(3) are satisfied, two data sets are indistinguishable in both integrated statistical-topological and biological sense.

Next, a dissimilarity between the original sample and subsample needs to be measured in one number. The following distance, delivering the dissimilarity between original sample S_0 and its reduced subsample S_n with corresponding cumulative distribution functions (CDF) F and G, is proposed:

$$D(S_0, S_n) = L_1(F, G) +$$
$$c_1 \cdot \mathbb{1}\{\dim(\vec{v}) \neq \dim(\vec{V})\} +$$
$$c_2 \cdot \sum_{i=1}^{dim(\vec{V})} \max(0, |v_i - V_i| - \varepsilon) \cdot \mathbb{1}\{\dim(\vec{v}) = \dim(\vec{V})\} +$$
$$c_3 \cdot \sum_{i=2}^{dim(\vec{V})} \max\left(0, \theta - \frac{|g(v_i) - g(v_{i-1}))|}{|f(V_i) - f(V_{i-1}))|}\right) \cdot \mathbb{1}\{\dim(\vec{v}) = \dim(\vec{V})\},$$

where

$F(l_j)$ and $G(l_j)$ are the CDFs values in length class l_j,

$L_1(F,G) = \sum_j |F(l_j) - G(l_j)|$ is a L_1-distance (also called 1-Wasserstein distance),

$f(l_j), g(l_j)$ are the LFD counts at the length class l_j,

$\vec{V} = sort(\vec{M}, \vec{A})$ and $\vec{v} = sort(\vec{m}, \vec{a})$ are the sorted increasing sequences of the robust modes and antimodes of the original sample and subsample, respectively,

$\mathbb{1}\{\Psi\}: \Psi \to \{0; 1\}$ is indicator function, i.e. $\mathbb{1}\{\Psi\} = 1$ if Ψ is true and $\mathbb{1}\{\Psi\} = 0$ otherwise,

c_1, c_2 and c_3 are some constants.

The first term is the L_1-distance between two CDFs as mentioned above, and the three next terms represent penalties, which are imposed for violation of constraints (1)-(3) in Definition 2. So, if a number of robust critical points in the subsample is different from this number in the original sample (violation of the condition (1)), then $\mathbb{1}\{\dim(\vec{v}) \neq \dim(\vec{V})\} = 1$ and the distance magnitude equals to $D = L_1(F,G) + c_1$, so a constant penalty is applied to infeasible LFD of subsample. In the same way, even by the equal number of modes/antimodes, the penalty term restrains their shifts: if the shift $|v_i - V_i|$ between some v_i and V_i exceeds ε (violation of the condition (2)), then $\max(0, |v_i - V_i| - \varepsilon) = |v_i - V_i| - \varepsilon$, and the distance magnitude increases, $D = L_1(F,G) + c_2 \cdot (|v_i - V_i| - \varepsilon)$. The constants c_1, c_2 and c_3 are introduced to define a hierarchy on constraints violation, although they can be put equal to 1.

Obviously, for $S_n \equiv S_0$ we obtain the lower bound $D = 0$. The upper bound can be provided by a minimally permitted reference subsample representing "the worst case". This is a reference subsample S_{ref}, which still reveals the patterns of the original distributional shape for given parameter values $(\Delta, \theta, \varepsilon)$ (i.e. meets conditions (1)-(3)), but cannot be reduced anymore because further subsampling will change the LFD shape. We will call the corresponding distance $D(S_0, S_{ref})$, where G^{ref} is the empirical CDF of S_{ref}, the admissible dissimilarity value (ADV). It represents a threshold (upper limit) to decide on acceptable and unacceptable dissimilarities between LFDs when reducing sampling effort. Thus, all subsamples $S = \{S_n\}$ with $D(S_0, S_n) \in [0; \text{ADV}]$ can be considered as representative ones in relation to the original target sample, thus, suitable to access the original length distribution information. It's easy to see that ADV= $D(S_0, S_{ref}) = L_1(F, G^{ref})$, since all penalty terms are equal to 0.

Note that the set of parameters $(\Delta, \theta, \varepsilon)$ that we apply for construction of the reference subsample, can be extended. One can introduce the following additional (optional) parameter γ, $\gamma < \theta$, which indicates a minimally required number per length class in a reference subsample. This parameter reflects the requirements of official national sampling programs in a certain sense (e.g. minimal fish number per length class needed for aging). Of course, the parameter γ can be also set to zero (i.e. ignored).

One can also consider only some part of length classes l^I (important length classes) for subsampling. For example, this can be just a middle part of the LFD, without large and small length classes.

Formally, the iterative algorithm scheme can be described as follows:

1) Use the standard RDB data with length rounded to 1 cm as basic input data.
2) Select bandwidth Δ and important length classes l^I if desired, identify corresponding robust modes/antimodes in the original sample under Δ on the set l^I.
3) Set remaining parameters $\{\theta, \gamma, \varepsilon\}$.
4) Remove one length measurement from each length class in l^I and see whether conditions (1)-(3) are satisfied. If yes, repeat the step. If no, go back to the previous subsample and stop. If a number of length measurements in some length class reaches value γ,

subsampling of this length class stops either, but subsampling of other length classes proceeds further until the conditions are met.

1.3.1.1 R-script

The corresponding generic iterative algorithm was implemented in the R-5.3.1 software. This R tool contains 4 functions, namely:

1) 01_set_RDB_Data: data manipulation function. The input is the required RDB tables (HL, CA, HH, SL). The function merges and filters the tables (e.g. selects species, years, areas etc.), and then transforms into a single table that's easier to work with for further analysis. As an example, a R-function translating the RDBES format into RDB is available at Annex 4. A graphical output contains the histogram of the result dataset.
2) 02_find_modes_antimodes: the function returns information on LFD structure of the sample, i.e. a list including:
    - All modes and antimodes;
    - Smoothed modes and antimodes for the desired bandwidth parameter Δ;
    - Robust modes and antimodes;
    - Vector of amplitudes.

   This function illustrates a setting of the Definition 1.
3) 03_minimal_reference_subsample_construction: the purpose of the function is a determining of the reference subsample. The input is the original sample (output of the function 1), robust modes and antimodes as well as amplitudes between them (output of the function 2), desired parameters set. The iteration procedure based on the Definition 2 produces a reference subsample as the function output. A graphical output is also incorporated, particularly, both histograms of the original sample and reference subsample.
4) 04_compute_distance: the function computes ADV between the original sample and reference subsample obtained as an output of the function 3. A graphical output displays a plot of both empirical CDFs of the original sample and reference subsample.

The case study presented in Section 4.1 demonstrates the application of the ADV-approach.

1.3.2 STREAMline (IB)

SDTool and BioSim tools were presented at WKBIOPTIM3 and a detailed description of both approaches, with some applications, can be found in the report (ICES 2019b).

SDTool was implemented for the first time in the MARE/2014/19 Med&BS project, and further improved as part of the STREAM project (MARE 2016/22). This tool allows to resample historical data, using bootstrap, for different stratifications (spatial, temporal, technical), with trips as the primary sampling unit. It produces a Coefficient of Variation (CV), raised LFDs, and the Earth Mover Distance (EMD) estimate.

BioSim Tool was implemented for the first time in the STREAM project (MARE 2016/22), based on the work carried out by WKBIOPTIM. This tool allows to resample historical data, using bootstrap, and to derive possible sub-samples of length measurements and an optimal number of individuals to be sampled for sex, maturity and age (the latter stratified by length class) by species. It produces a Coefficient of Variation (CV) and the Earth Mover Distance (EMD) estimate.

New developments will take place in the STREAMline project (MARE/2020/08):

1. Development of additional quality indicators to the ones developed and tested in STREAM taking into account:

2. the work carried out in ICES WKBIOPTIM3
3. the Admissible Dissimilarity Value (ADV), as a measure of sampling reliability (Wischnewski et al., 2020)
4. Evaluation of the variability of relevant estimates (e.g. von Bertalanffy parameters, size at first maturity) and identification of a satisfactory sub-sampling strategy;
5. Development of SDTool and BioSim Tool to allow the extraction of samples from the dataset used for the case study, according to combinations of technical, time and spatial characteristics that could be relevant for specific case studies in order to draft a Regional Work Plan (e.g. Country-Geographical Sub-Area, sensu GFCM). The sampling design could be set according to a sampling hierarchy (e.g. from individual fish to trip) indicating what sampling levels are included in the multi-stage sampling of the commercial catches and how they are (hierarchically) related to each other, in line with the RDBES concept.

1.3.3 SampleOptim (PG)

SampleOptim has been designed and implemented based on the Portuguese National Programme for Biological Sampling (EU Data Collection Framework). The main objective is to determine the optimal number of fish per length class to sample in order to produce input data (e.g. ALKs – age length keys and MO – maturity ogive) for species stock assessment. SampleOptim's main feature is the ability to perform simulations allowing the consideration of some stratification conditions based on: temporal (annual, semester or quarter), ports (uniformly selection of ports/randomly ports selection or gears) and sexratio (setup the sexratio of subsample).

The simulation process works at the sample level. It relies on a dataset that represents the "whole" population, and the simulations are based on randomly subsampling without replacement (although there is a built-in option for the user to choose to sample with replacement - see on Section 4.2. an example comparing the results from applying those two types of subsampling). ALKs and MOs estimates are produced and compared, based on a reduction of the number of individuals sampled by length class. However, in cases where the original sample size is not enough, the functions in the algorithms will not converge, which could indicate some bias in the original data sampling.

A series of quality indicators are estimated to evaluate the different scenario tested and to help the decision process of finding the optimum sample size by length class, accounting that ALKs and MOs are of interest for stock assessment purposes. The following variables are produced to compare the original versus simulated datasets: mean length-at-age, mean age-at-length, coefficient of variation length-at-age, coefficient of variation age-at-length, standard deviation length-at-age, standard deviation age-at-length, the parameters of the von Bertalanffy growth model (Linf, t0 and k) and the maturity ogive parameters (L25, L50 and L75). Besides the latter quality indicators, the root mean squared prediction error (RMSPE), mean squared prediction error (MSPE or MSE) and mean average percentage error (MAPE) are also calculated.

SampleOptim was presented at WKBIOPTIM3 (see for details: ICES, 2019 - Section 2.3 and Annex 3A). An outline of the approach is presented in Figure 1.4.3.

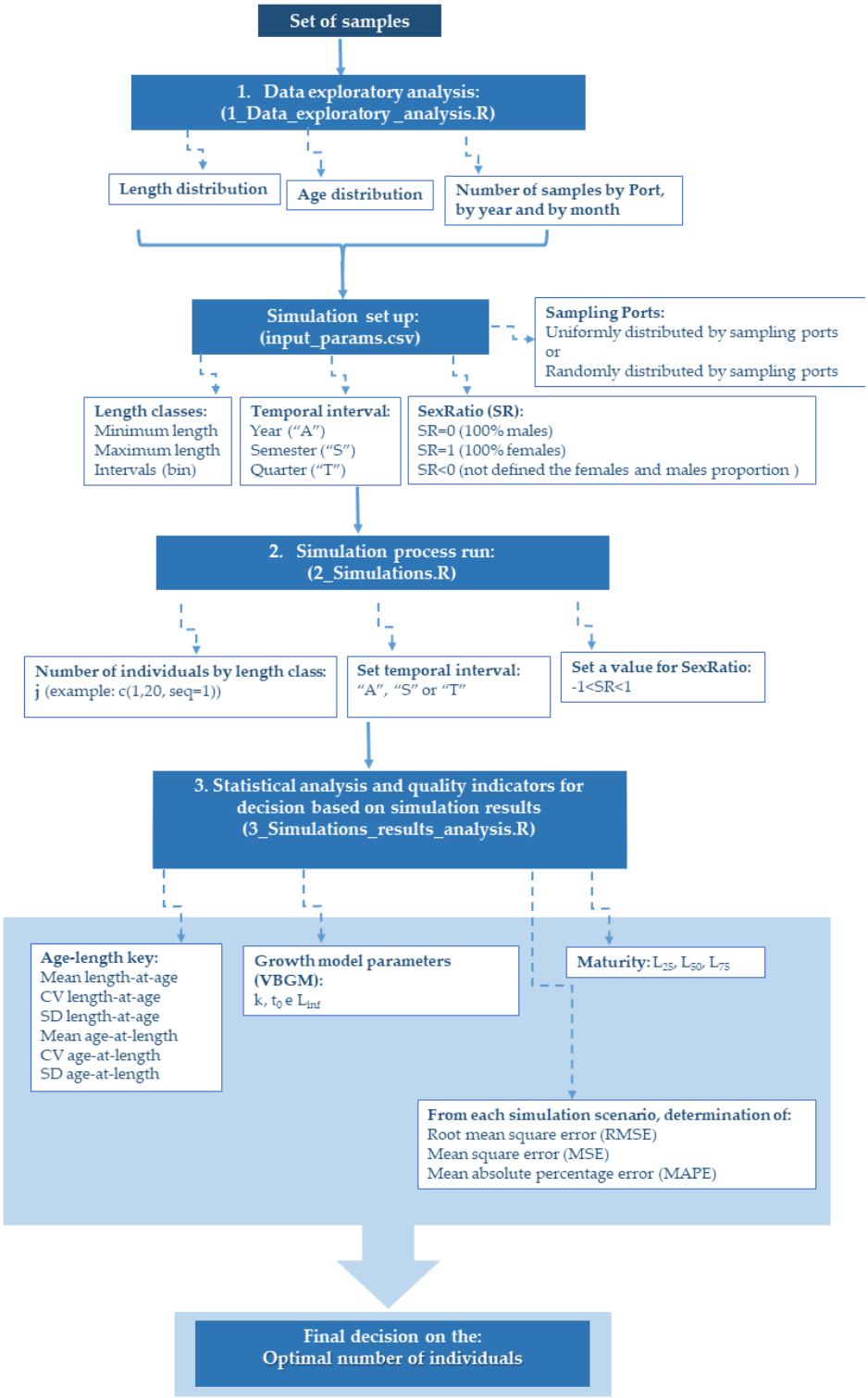


Figure 1.4.3. SampleOptim flowchart showing the optimization procedure with the indication of the different R scripts, steps, type of options to setup the subsamples simulations and the different quality indicators used to assist the user with making a decision in the optimum sample size.

1.3.4 FishPi2 (GL)

The R package of FishPi[2] WP3, which aimed at simulating length sampling, was initially developed based on the main simulation framework of the FishPi project, which aimed at testing regional sampling designs to estimate landings. Some adaptations were made during the third workshop WKBIOPTIM3 to produce simulation outputs comparable with the other approaches presented. These were all included in an R package called FishPi4WKBIOPTIM. At this stage, no further developments have been made on this package which remains available to the group and which still has great potential to be built upon in line with future directions of work in the field.

1.4 Quality indicators

Each tool produces some descriptive and inferential statistics, referred to as indicators. See summary of the WKBIOPTIM3 case studies that aimed at comparing tool outcomes (Table 1.5.1) and corresponding indicator overviews (Table 1.5.2).

The application of different tools using the same input data, in the WKBIOPTIM3 case studies, highlighted the fact that the outputs of the different approaches were not easily comparable. In some cases, the statistical outputs could even be challenging to interpret for users not involved in developing the methods. This issue was addressed during WKBIOPTIM4 with the premises of a common R package. Table 1.5.3 provides information on the data requirements for each indicator, what they are used for, and which other tool might be able to produce it.

Table 1.5.1. List of case studies from WKBIOTPIM 3, main tools applied and outputs produced.

Case study	Tool 1	Tool 2	Indicators
Plaice (ICES Div. 27.4.b) Haddock (ICES Div. 27.7.e) Plaice (ICES Div. 27.7.d)	SampleReferenceLevel	-	Admissible Dissimilarity Value (ADV)
Greece purse seine fleet	WKBIOPTIM2 Multi-level analysis	SDTool	Weighted CV (WCV)
***Mullus barbatus* (Southern Adriatic Sea)**	SampleOptim	BioSim Tool	SampleOptim: Mean absolute percentage error (MAPE), CV BiomSim Tool: CV
***Mullus barbatus* (Southern Adriatic Sea)**	SDTool	FishPi4WKBIOPTIM	SDTool: CV FishPi4: Relative Standard Error (RSE), mean length
Sandeel (greater North Sea)	FishPi4WKBIOPTIM	SDTool	SDTool: CV FishPi4: RSE, mean weigthed (MWCV)
***Mullus barbatus* (Aegean Sea- Greece)**	SampleOptim	BioSim Tool	SampleOptim: MAPE, Mean Squared Prediction Error (MSPE), Root Mean Squared Prediction Error (RMSPE), CV. BioSim Tool: CV

Table 1.5.2. Indicator, type of statistics produced, description of what it does, and the corresponding tool that originally produces it

Indicator	Type	What it does	Which tool it is originally an output of
mean, median, minimum and maximum length	Descriptive	Describes some aspects of the simulated sampled length distribution	BioSimTool
Mean Weighted CV (MWCV)	Descriptive	Provides a description of the precision over the entire range of a length frequency distribution	BioSimTool
Earth Mover's Distance (EMD)	Inferential	Provides an estimate of the similarity between 2 distributions; the sample and the population	BioSimTool SDTool (function from R package *emdist*)
CV sex ratio at length, maturity at length, ALK	Descriptive	Description of precision	BioSimTool
CV per length class and total	Descriptive	Provides a description of the precision over the entire range of a length frequency distribution taking into account the sampling stratification.	SDTool – function from the COST library
mean length-at-age, mean age-at-length	Descriptive	Describes some aspects of the simulated sampled length at age distribution	SampleOptim
parameters of the VB growth model	Descriptive	Provides an estimate of the parameters from the von Bertalanffy growth model (t0, k, Linf).	SampleOptim
maturity ogive parameters	Descriptive	Provides estimates of maturity at length, mainly: L25, L50 and L75	SampleOptim
Root Mean Squared Prediction Error (RMSPE)	Inferential	Provides an estimate of the standard deviation of the residuals (prediction errors).	SampleOptim
Mean Squared Prediction Error (MSPE)	Inferential	Provides an estimate of the standard deviation of the residuals (prediction errors).	SampleOptim
Mean Average Percentage Error (MAPE)	Inferential	Provides an estimate of the standard deviation of the residuals (prediction errors).	SampleOptim
Admissible Dissimilarity Value (ADV)	Inferential	Estimate of dissimilarity between original and sampled length distribution	SampleReferenceLevel
Relative Standard Error (RSE)	Descriptive	Calculated as the standard deviation over the mean of estimates (across simulation replicates)	FishPi4WKBIOPTIM

Table 1.5.3. Information on data requirements and objectives of the indicators, and what other tool could produce it.

Indicator	Data required	Meaning/objective	Which R-tool can it be applied to
mean, median, minimum and maximum length	Length data of sample and population	Compare with the original population sampled used as a reference level.	Any tool
Mean Weighted CV (MWCV)	Length data of sample and population	The smaller the CV the better.	Any tool
Earth Mover's Distance (EMD)	Length data of sample and population	The smaller the EMD the better. EMD measure the distance between two probability distributions (e. g. sample and population)	Any tool
CV sex ratio at length, maturity at length, ALK	Sex, maturity, age-length data of sample and population.	The smaller the CV the better. The precision of the sample for sex, maturity and age is evaluated versus the total of individuals sampled for the length, by length class. ALKs from the subsample are used to compare with the original population sampled (used as a reference). The precision is evaluated by the age length distribution obtained on the subsamples from the simulations.	BioSim, SampleOptim
CV per length class and total	Length frequency distribution of sample and population	The smaller the CV the better.	Any tool
mean length-at-age, mean age-at-length	Lengths and ages of sample and population.	Compare with the original population sampled used as a reference level (population).	BioSim, SampleOptim
parameters of the VB growth model	Lengths and ages of sample and population.	Compare with the original population sampled used as a reference level.	BioSim, SampleOptim
maturity ogive parameters	Maturity and lengths of sample and population.	Compare with the original population sampled used as a reference level.	BioSim, SampleOptim
Root Mean Squared Prediction Error (RMSPE)		Used as a measure of precision to decide on the optimum sample size.	Any tool
Mean Squared Prediction Error (MSPE)		Used as a measure of precision to decide on the optimum sample size.	Any tool
Mean Average Percentage Error (MAPE)		Used as a measure of precision to decide on the optimum sample size.	Any tool
Admissible Dissimilarity Value (ADV)	Length data of sample and population		Any tool
Relative Standard Error (RSE)			Any tool

1.5 R-package

1.5.1 Process

An R-package has been created and the source code is hosted on GitHub at:

https://github.com/ices-eg/WKBIOPTIM4

Hosting in this way allows the tools to be easily available, well documented, and compatible, able to be installed on Mac/PC/Linux OS's and any version of R.

GL and NC have acted as the maintainers. GitHub allows the use of separate 'branches' which has enabled other contributors to work on adding tools, data, documentation and functions independently before these are merged to a main branch. Use of the 'roxygen2' package has streamlined the adding of documentation and examples.

1.5.2 Functions description

During the workshop, the developers' sub-group started to code the functions for the indicators. Table 1.6.2 reports the state of completion of these functions, as off the end of the workshop week. Four functions were written up to be added to the GitHub repository: MWCV, maturity ogive parameters, CV of ALK and summary indicators. The description of those functions and the type of the input data needed for each function is presented in sections 1.6.2.1 to 1.6.3.4.

Table 1.6.2. State of completion of the indicators' functions during the workshop.

Indicator	Status
MWCV	√
Earth Mover's Distance (EMD)	emd2 package
CV per length class and total	
Mean age-at-length	
Parameters of the von Bertalanffy growth model	
Maturity ogive parameters	√
Mean squared prediction error (MSPE)	
Mean average percentage error (MAPE)	
Admissible Dissimilarity Value	
Relative standard error (RSE)	FishPi4WKBIOPTIM package
CV of ALK by age and total	√
Summary statistics (mean length, median, se, min, max, number of sampled classes)	√

1.5.2.1 MWCV (mean weighted CV) calculation

MWCV {BIOPTIMtools} R Documentation

Description

MWCV (mean weighted CV) calculation

Usage

MWCV(df1, variable)

Arguments

df1: data frame of sampled data by length class in CA format (RDB) (individual measurements)

variable: name of the column containing the length measurements (as character)

Value

MWCV

Examples

MWCV(example_samples,"lenCls")

1.5.2.2 Maturity ogive parameters estimation (L25, L50, 75)

Maturity_ogive {BIOPTIMtools} R Documentation

Description

Maturity ogive parameters estimation (L25, L50, L75)

Usage

Maturity_ogive(data)

Arguments

Data: Dataframe with information from the simulations, containing the following variables: length, maturity (0- immature; 1 - mature); IDsim (identification of the number of the simulation run); type (number of individuals selected in the current simulation, e.g. 10, 20, 30, ...).

Value

L25 (length at which 25 percent of the individuals are mature)

L50 (length at which 50 percent of the individuals are mature)

L75 (length at which 75 percent of the individuals are mature)

1.5.2.3 Age-Length Key CV (by age and total)

CV_ALK {BIOPTIMtools} R Documentation

Description

Function to calculate CV on ALK

Usage

CV_ALK(DF)

Arguments

DF: dataframe with lengths in the first column and ages on the other columns

Value

CV by age class and total CV

Examples

CV_ALK(example_ageLength)

1.5.2.4 Make_summary_numeric

Description

Summary statistics calculation: mean length, se, median, min, max, n classes sampled

Usage

make_summary_numeric(df1, variable, a, b)

Arguments

df1: data frame of sampling data

variable: "lenCls"

a: coefficient of length-weight relationship

b: coefficient of length-weight relationship

Value

table reporting the different estimates

Examples

make_summary_numeric(example_samples,"lenCls",a=0.0006,b=3)

2 RDBES

2.1 Brief overview of RDBES data model

In this section, the text was adapted and summarized from "Documentation of the Regional Database and Estimation System - RDBES Data Model doc. v. 1.19.2 (23 September 2021)".

RDBES is the new Regional Database and Estimation System that is currently being developed by ICES and, in the longer term, it will replace both the current RDB and InterCatch systems, providing a single platform for countries to produce statistical estimates of quantities of interest, to be used as inputs for the assessment groups.

The aims of the RDBES are:

1) To ensure that data can be made available for the coordination of regional fisheries data sampling plans, in particular for the EU DC-MAP Regional Coordination Groups (RCGs);

2) To provide a regional estimation system such that statistical estimates of quantities of interest can be produced from sample data;

3) To increase the data quality, documentation of data and ensuring of approved estimation methods are used;

4) To serve and facilitate the production of fisheries management advice and status reports;

5) To increase the awareness of fisheries data collected by the users of the RDBES and the overall usage of these data.

The RDBES data model allows the accommodation of different designs present in the national sampling programmes and it includes a number of different hierarchies, representing the different sampling techniques that are used in practice. Two categories of hierarchies are used in the model - the upper hierarchy describes how a sample is selected, and the lower hierarchy that describes what type of length-frequency or biological variables are measured for that sample.

The **RDBES documentation** can be found in: https://github.com/ices-tools-dev/RDBES/blob/master/Documents/RDBES%20Documentation%20of%20the%20Data%20Model.docx

- Hierarchies, general: p. 14
- Hierarchies, detailed: Annex 1, p. 28
- Stratification: Section 'Stratification ', p. 19
- Commercial Landings (CL) and Commercial Effort (CE): Section 'Aggregated Commercial Landings (CL) and Commercial Effort (CE)', p. 7

The **RDBES data model links:**

- Commercial Landings (CL) and Commercial Effort (CE): https://github.com/ices-tools-dev/RDBES/blob/master/Documents/RDBES%20Data%20Model%20CL%20CE.xlsx
- Commercial Sampling (CS): https://github.com/ices-tools-dev/RDBES/blob/master/Documents/RDBES%20Documentation%20of%20the%20Data%20Model.docx &

https://github.com/ices-tools-dev/RDBES/blob/master/Documents/RDBES%20Data%20Model%20VD%20SL.xlsx

a) More comprehensive description of the lower hierarchies

i. Lower hierarchies include four types of collecting biological samples and they all link to the Upper Hierarchies with the SA table (sample). The tables included in these hierarchies are the frequency measure (FM) and biological variable (BV).

1. Lower Hierarchy A: Length stratified biological samples - both FM and BV tables are present
2. Lower Hierarchy B: Only length frequency data is taken from sample(s)/subsample(s) - only FM table is present
3. Lower Hierarchy C: All individuals in the sample/subsample are biologically analysed - only BV table is present
4. Lower Hierarchy D: No length measurements or biological analyses - no tables present

b) Brief description of the main differences between the two formats:

There are major differences for the Commercial Sampling (CS) information between the two formats. In RDB, there are five common tables (TR (trip), HH (station/haul), SL (species list), HL (haul length) and CA (catch aged)) that are used by countries for all types of commercial sampling data (onboard, onshore, and biological sampling). In the RDBES the structure is more complex and the number and kind of tables present is dependent on the type of sampling, and also accommodates the sampling design (hierarchy) used by each country. This means that the information present in one table from the RDB can be present in different types of tables from the RDBES, according to the hierarchy adopted. An example for compiling the information present in the CA table in RDB using the new RDBES format is presented in Table 2.1.1 (upper hierarchies) and Table 2.1.2 (lower hierarchies). The complexity presented in these tables showed how difficult would be to convert the RDB to the RDBES format.

Table 2.1.1. Example of an attempt to convert the CA table variables from RDB into the RDBES format, accounting for the upper hierarchies.

RDB		RDBES Upper hierarchies					
RDBTable	RDBVariable	Hierarchy 1		Hierarchy 2		Hierarchy 3	
		RDBESTable	RDBESVariable	RDBESTable	RDBESVariable	RDBESTable	RDBESVariable
CSCA	CS_TripId	FT	FTid	FT	FTid	FT	FTid
CSCA	SamplingType	FT	FTsamplingType	FT	FTsamplingType	FT	FTsamplingType
CSCA	LandingCountry	FT	[FTarrivalLocation]	FT	[FTarrivalLocation]	FT	[FTarrivalLocation]
CSCA	VesselFlagCountry	VD	VDflagCountry (link to VD from VS | FT)	VD	VDflagCountry (link to VD from FT)	VD	VDflagCountry (link to VD from VS | FT)
CSCA	Year	DE	DEyear	DE	DEyear	DE	DEyear
CSCA	Quarter	FT, FO	[FTarrivalDate], [FOstartDate], [FOendDate]	FT, FO	[FTarrivalDate], [FOstartDate], [FOendDate]	FT, FO	[FTarrivalDate], [FOstartDate], [FOendDate]
CSCA	Month						
CSCA	Project	DE	DEsamplingScheme	DE	DEsamplingScheme	DE	DEsamplingScheme
CSCA	Trip	FT	FTsequenceNumber	FT	FTsequenceNumber	FT	FTsequenceNumber
CSCA	StationNo	FO, OS, LE	FOsequenceNumber, OSsequenceNumber,	FO	FOsequenceNumber	FO	FOsequenceNumber
CSCA	Species	SA	SAspeciesCode, (SAspeciesCodeFAO)				
CSCA	Sex	SA	SAsex				
CSCA	CatchCategory	SA	SAcatchCategory				
CSCA	LandingCategory	SA	SAlandingCategory				
CSCA	SizeCategoryScale	SA	SAcommSizeCatScale				
CSCA	SizeCategory	SA	SAcommSizeCat				
CSCA	Area	SA	SAarea				
CSCA	StatisticalRectangle	SA	SArectangle				
CSCA	Subpolygon	SA	SAgsaSubarea				

[information can be extracted from this variable]
(optional table/field)
Information from SA table is common to all hierarchies
No information

Table 2.1.1. (continued). Example of an attempt to convert the CA table variables from RDB into the RDBES format, accounting for the upper hierarchies.

RDB		RDBES Upper hierarchies					
RDBTable	RDBVariable	Hierarchy 4		Hierarchy 5		Hierarchy 6	
		RDBESTable	RDBESVariable	RDBESTable	RDBESVariable	RDBESTable	RDBESVariable
CSCA	CS_TripId	FT	FTid	(FT)	(Ftid)	FT	FTid
CSCA	SamplingType	FT	FTsamplingType	(FT)	(FTsamplingType)	FT	FTsamplingType
CSCA	LandingCountry	LE	LEcountry	LE	LEcountry	FT	[FTarrivalLocation]
CSCA	VesselFlagCountry	VD	VDflagCountry (link to VD from FT | LE)	VD	VDflagCountry (link to VD from LE)	VD	VDflagCountry (link to VD from FT)
CSCA	Year	DE	DEyear	DE	DEyear	DE	DEyear
CSCA	Quarter	OS, FT, LE	[OSsamplingDate], [FTdepartureDate], [FTarrivalDate], [LEdate]	OS, (FT), LE	[OSsamplingDate], [FTdepartureDate], [FTarrivalDate], [LEdate]	OS, FT, (LE)	[OSsamplingDate], [FTdepartureDate], [FTarrivalDate], [LEdate]
CSCA	Month						
CSCA	Project	DE	DEsamplingScheme	DE	DEsamplingScheme	DE	DEsamplingScheme
CSCA	Trip	FT	FTsequenceNumber	(FT)	(FTsequenceNumber)	FT	FTsequenceNumber
CSCA	StationNo	OS, LE	OSsequenceNumber, LEsequenceNumber	OS, LE	OSsequenceNumber, LEsequenceNumber	OS, FO	OSsequenceNumber, FOsequenceNumber
CSCA	Species						
CSCA	Sex						
CSCA	CatchCategory						
CSCA	LandingCategory						
CSCA	SizeCategoryScale						
CSCA	SizeCategory						
CSCA	Area						
CSCA	StatisticalRectangle						
CSCA	Subpolygon						

[information can be extracted from this variable]
(optional table/field)
Information from SA table is common to all hierarchies
No information

Table 2.1.1. (continued). Example of an attempt to convert the CA table variables from RDB into the RDBES format, accounting for the upper hierarchies.

RDB		RDBES Upper hierarchies					
RDBTable	RDBVariable	Hierarchy 7		Hierarchy 8		Hierarchy 9	
		RDBESTable	RDBESVariable	RDBESTable	RDBESVariable	RDBESTable	RDBESVariable
CSCA	CS_TripId			(FT)	Ftid		
CSCA	SamplingType			(FT)	FTsamplingType		
CSCA	LandingCountry	OS	[OSlocode]	(FT)	[FTarrivalLocation]	LO	[Lolocode]
CSCA	VesselFlagCountry	VD	VDflagCountry (link to VD from LE)	VD	VDflagCountry (link to VD from VS | LE)	VD	VDflagCountry (link to VD from LE)
CSCA	Year	DE	DEyear	DE	DEyear	DE	DEyear
CSCA	Quarter	OS, (LE)	[OSsamplingDate], [LEdate]	LE, (FT)	[LEdate], [FTdepartureDate], [FTarrivalDate]	(LE)	[LEdate]
CSCA	Month						
CSCA	Project	DE	DEsamplingScheme	DE	DEsamplingScheme	DE	DEsamplingScheme
CSCA	Trip			FT	FTsequenceNumber		
CSCA	StationNo	OS, (LE)	OSsequenceNumber, (LEsequenceNumber)	LE	LEsequenceNumber	LO, TE	LOsequenceNumber, TEsequenceNumber
CSCA	Species						
CSCA	Sex						
CSCA	CatchCategory						
CSCA	LandingCategory						
CSCA	SizeCategoryScale						
CSCA	SizeCategory						
CSCA	Area						
CSCA	StatisticalRectangle						
CSCA	Subpolygon						

[information can be extracted from this variable]
(optional table/field)
Information from SA table is common to all hierarchies
No information

Table 2.1.1. (continued). Example of an attempt to convert the CA table variables from RDB into the RDBES format, accounting for the upper hierarchies.

RDB		RDBES Upper hierarchies							
RDBTable	RDBVariable	Hierarchy 10		Hierarchy 11		Hierarchy 12		Hierarchy 13	
		RDBESTable	RDBESVariable	RDBESTable	RDBESVariable	RDBESTable	RDBESVariable	RDBESTable	RDBESVariable
CSCA	CS_TripId	FT	FTid	FT	FTid	(FT)	Ftid	(FT)	Ftid
CSCA	SamplingType	FT	FTsamplingType	FT	FTsamplingType	(FT)	FTsamplingType	(FT)	FTsamplingType
CSCA	LandingCountry	FT	[FTarrivalLocation]	LO, FT	[LOlocode], [FTarrivalLocation]	LO, (FT)	[LOlocode], [FTarrivalLocation]	(FT)	[FTarrivalLocation]
CSCA	VesselFlagCountry	VD	VDflagCountry (link to VD from VS | FT)	VD	VDflagCountry (link to VD from FT | LE)	VD	VDflagCountry (link to VD from LE)		
CSCA	Year	DE	DEyear	DE	DEyear	DE	DEyear	DE	DEyear
CSCA	Quarter	FT, FO	[FTdepartureDate], [FTarrivalDate], [FOstartDate], [FOendDate]	FT, (LE)	[FTdepartureDate], [FtarrivalDate], [LEdate]	LE, (FT)	[LEdate], [FTdepartureDate], [FtarrivalDate]	FO, (FT)	[FOstartDate], [FOendDate], [FTdepartureDate], [FTarrivalDate]
CSCA	Month								
CSCA	Project	DE	DEsamplingScheme	DE	DEsamplingScheme	DE	DEsamplingScheme	DE	DEsamplingScheme
CSCA	Trip	FT	FTsequenceNumber	FT	FTsequenceNumber	(FT)	FTsequenceNumber	(FT)	FTsequenceNumber
CSCA	StationNo	FO	FOsequenceNumber	LE	LEsequenceNumber	LE	LEsequenceNumber	FO	FOsequenceNumber
CSCA	Species								
CSCA	Sex								
CSCA	CatchCategory								
CSCA	LandingCategory								
CSCA	SizeCategoryScale								
CSCA	SizeCategory								
CSCA	Area								
CSCA	StatisticalRectangle								
CSCA	Subpolygon								

[information can be extracted from this variable]
(optional table/field)
Information from SA table is common to all hierarchies
No information

Table 2.1.2. Example of an attempt to convert the CA table variables from RDB into the RDBES format, accounting for the lower hierarchies.

RDB		RDBES Lower hierarchies							
RDBTable	RDBVariable	Hierarchy A		Hierarchy B		Hierarchy C		Hierarchy D	
		RDBESTable	RDBESVariable	RDBESTable	RDBESVariable	RDBESTable	RDBESVariable	RDBESTable	RDBESVariable
CSCA	Species	[SA]	SAspeciesCode, (SAspeciesCodeFAO)	[SA]	SAspeciesCode, (SAspeciesCodeFAO)	[SA]	SAspeciesCode, (SAspeciesCodeFAO)	No data	
CSCA	Sex	BV	BVtypeMeasured::Sex	No data		BV	BVtypeMeasured::Sex		
CSCA	LengthCode	BV	Bvaccuracy			BV	Bvaccuracy		
CSCA	AgingMethod	BV	BVmethod			BV	BVmethod		
CSCA	LengthClass	FM	FMclassMeasured	FM	FMclassMeasured	BV	BVtypeMeasured::LengthTotal (+ other type length measurements)		
CSCA	Age	BV	BVtypeMeasured::Age	No data		BV	BVtypeMeasured::Age		
CSCA	AgePlusgroup								
CSCA	OtolithWeight	BV	BVtypeMeasured::InfoOtolithMorphometrics			BV	BVtypeMeasured::InfoOtolithMorphometrics		
CSCA	OtolithSide	BV	BVtypeMeasured::InfoOtolithMorphometrics			BV	BVtypeMeasured::InfoOtolithMorphometrics		
CSCA	Weight	BV	BVtypeMeasured::WeightMeasured			BV	BVtypeMeasured::WeightMeasured		
CSCA	MaturityStagingMethod	BV	BVmeasurementEquipment			BV	BVmeasurementEquipment		
CSCA	MaturityScale	BV	BVvalueUnitOrScale			BV	BVvalueUnitOrScale		
CSCA	MaturityStage	BV	BVtypeMeasured::Maturity			BV	BVtypeMeasured::Maturity		
CSCA	SingleFishId	BV	BVfishId			BV	BVfishId		

[information can be axtracted from this variable]
(optional table/field)
:: - "droplist" selection for type of variable
Not present in RDBES

c) Changes to the Commercial Landings (CL) and Effort (CE)

1. The CL and CE tables are to collect aggregated data from national commercial fisheries. Both tables have the same structure between the two formats but the one from RDBES presents more detailed information namely it allows for both official and scientific estimates of landings and effort to be presented, and also includes uncertainty indicators (RSE or qualitative bias) when estimates are provided.

2.2 Transition to RDBES model

The importance of using the RDBES data in these tools mainly relates to the benefits in considering the sampling design in the optimization procedures, because the data provided by RDBES is more complete and detailed (e.g. different types of measures), also aiming to reproduce the exact way the sampling was performed.

The translation of the RDB format related tools into the RDBES will become a need also because it is expected that historical data will be reported in the new RDBES format. However, that translation is not very easy and straightforward to accomplish at this stage, especially for the Sampling Level R-tools, that will need an adaptation to include/accommodate the type of sampling design (Upper Hierarchies) adopted. This transition to the new format will not be easy for some of those tools and, probably, can only be developed after the estimation procedures for the RDBES (WKRDB-EST2) ICES, 2022) are implemented, and by combining efforts between the two groups (estimation and optimization). Regarding the Sample Level R-tools, it will probably be more easy to adapt because they're only accounting for information from the Lower Hierarchies but, anyway, the code will still need to be adjusted according to the Lower Hierarchy considered.

2.3 Linkage of the different hierarchies to the R-tools

The group discussed how the different hierarchies could be related to each of the WKBIOPTIM tools. For some of the tools the upper hierarchy can be ignored (e.g. BioSim, SampleOptim and SampleLevelOptim) but not the lower one (A, B and C for BioSim, B for SampleOptim and SampleLevelOptim). For others, it was possible to identify a category candidate for the upper hierarchy to use (e.g upper hierarchies H1 for SampleReferenceLevel, H2 for SDTool and SimPop and H1, H2, H4 for Fishpi4WKBIOPTIM; and H12 for the sample collection on SampleOptim). Table 2.3 presents a summary of the WKBIOPTIM R-tools, including some additional details on the RDBES hierarchies (lower and upper) "assignment".

In the table below the main characteristics and features of the R-tools developed and/ or applied under WKBIOPTIM are presented (Table 2.3).

Table 2.3. WKBIOPTIM tools main characteristics and features.

	R-toools	SDtool	BioSim Tool	Fishpi4WKBIOPTIM	SimPop	SampleOptim	SampleLevelOptim	SampleReferenceLevel
	R_developers	Isabella Bitetto	Isabella Bitetto	Glwadys Lambert	Laurent Dubroca	Patrícia Gonçalves	Nuno Prista	Julia Wischnewski
VARIABLE(S) OF INTEREST (what the precision is estimated for)	Landings (or discards or catch)	No	No	No	No	No	No	No
	Mean length	No	No	Yes	No	No	yes	No
	Length distribution	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Mean age at length	No	No	No	No	Yes	No, but can be extended	No
	Age distribution	No	Yes	No	No	Yes	Yes	No, but can be extended
	Sex ratio	No	Yes	No	No	Yes	Yes	No
	Maturity at age	No	Yes	No	No	Yes	Yes	No
	Inference on sample [1] or population (raised within the tool) [2] data	1	1	1	2	2	1	2
SAMPLING LEVELS	PSU/SSU/TSQ/QSU (examples)	# Trip ~~# Trip/haul~~	Fish	# Trip # vessel/Trip # site-day/ Trip	Trip/haul	Fish	Fish - picks samples with a sample e.g. a fishing trip - lower hierachy B and C, a box of fish. How this works with length stratified vs. none	Trip, vessel, metier
	RDBES upper hierachies	# Trip - H2 with hauls (FO) aggregated at the trip level	Ignored	H2 with hauls (FO) aggregated at the trip level	H2	Ignored - picking from a pool with all fish sampled (the samples used in development are coming from hierachy 12)	Not relevant - resampling within a sample and results at the level of the sample	H1
				H1 with hauls (FO) aggregated at the trip level				
		# Trip/haul - ?		H4				
	If the upper hierachy is ignored in the resampling is it then done within a sampling unit present e.g. haul, landing event		Fish within a original sample for length. For other biological measurements are picking from a pool of all fish					
	RDBES lower hierachies	Need to find the length, but need to take A (only using the FM), B and C into account	A B, C	Not relevant - All length frequencies are raised to the trip level in input data. The function translating the RDB data into the input will require some work	A, B and C	A (only using the BV)	B (random sample of fish)	
	Requires length stratified of biological raised to LD?	Not relevant	?	Not relevant	Not relevant		Assumes random sample of biological measurements	Not relevant, when only lengths
	Handling of length stratified sampling of other biological measurements. In the present RDB this is not apparent in the data, but it is possible to check	Not relevant		Not relevant	Not relevant	how is the length stratification in the original data handled?		

Table 2.3. (continued). WKBIOPTIM tools main characteristics and features.

	R-toools	SDtool	BioSim Tool	Fishpi4WKBIOPTIM	SimPop	SampleOptim	SampleLevelOptim	SampleReferenceLevel
SAMPLING HIERARCHY	1: Single level unstratified	1, 2	1	1, 2	3	1,2	1, 2	1
	2: Single level stratified							
	3: Multi-level							
STRATIFICATION OPTIONS	Technical (eg metier)	Yes	N/A	Yes	Yes	Yes (by selecting specific metiers)	yes (choice of specific metiers)	yes (choice of specific metiers)
	Spatial (eg GSA, ICES areas)	Yes	N/A	Yes	Yes	Yes (by selecting specific an ICES area, or GSA)	yes (choice of specific ICES area)	yes (choice of specific ICES area)
	Temporal (eg year)	Yes	N/A	No	Yes	Yes (by selecting specific quarter, semester, year)	yes (choice of specific year and quarter)	yes (choice of specific year and quarter)
	other (specify)	N/A	N/A	N/A	Vessel paramters (length, GT...)	Sexratio, port	any biological variable (e.g, length, sex)	choice of specific sampling type
	RDBES (free text for all sampling unit levels)	Fixed options	N/A	Stratification is free text at the PSU level, but the input need to be in the input data - to be checked	Stratification is free text at the PSU level, but the input need to be in the input data	Fixed	Not relevant - resampling within a single sample	Stratification is free text at the PSU level
CONCURRENT SAMPLING	1: all species in sample	3	3	3	1 & 4	4	4	4
	2: can select which species to sample							
	3: can select mulitple species but will run one by one							
	4: one species at a time							
INPUT DATA	RDB	RDB (CS, CA, TR, HH, HL, CL)	RDB (CA, HH)	RDB (HL, HH, SL, TR)	RDB (CS, CA, TR, HH, HL, CL)	RDB (CA, HH)	Yes (CA, HH)	RDB: HL, HH, HL (for length). Can be added: CA (for age), CL (for raising of LFD to metier level)
	Other format	Cost objects (CS and CL)	N/A	FishPi data call	No	N/A	Datras, HH (after conversion to CA format)	Datras

Table 2.3. (continued). WKBIOPTIM tools main characteristics and features.

	R-toools	SDtool	BioSim Tool	Fishpi4WKBIOPTIM	SimPop	SampleOptim	SampleLevelOptim	SampleReferenceLevel
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)
	Example_Question	Would the precision of the sampling have been very different if we had sampled more trips mantaining the same number of individuals measured?	Would the precision of the sampling and the derived biological information (sex ratio, maturity ogive, age structure) have been very different if we had sampled less individuals per sample?	What would be the precision in length estimations for a given species with a given sampling design (with a given choice of strata and sampling effort at the trip level) for a domain of interest (e.g. at the stock level)?	What is the effect of decrease the number of samples on population estimates?	Would the precision of the sampling and the derived biological information (maturity ogive, age-at-length) have been very different if we had sampled less individuals (e.g. per sample, quarter, port, metier/fleet)?	Would the results of a sample (or set of samples) have been very different if we had sampled less individuals?	Would it be possible to reduce the number of metiers/trips/hauls/measured individuals in area/domain X without significant affecting the LFD for species Y? The underlying LFD depends on chosen aggregation level (raising at metier level, trip level etc.) and is defined by practical goals (e.g. we compare our national LFD for species Y in area X with corresponding LFD of all EU states).

2.3.1 Stratification

In the RDB format, no information on stratification is available at any of the sampling levels. In the RDBES it is possible to declare stratification on all the sampling levels. The field for stratification is a free text field, so it is possible to support the variety of stratification being used in commercial catch sampling e.g. size sorting categories, distance to sampling facilities and big/small harbours.

Most of the tools in the WKBIOPTIM suite have a fix set of possible stratification e.g. time, space and gear selectivity, which do not always fit the stratification needed in real sampling programs, so it would be beneficial to make the stratification options more generic in the future e.g. free text.

Further, when using past samples for optimization it is important to take the stratification used in the past into account, e.g. it is tricky to combine age samples from random sampling with age samples from length stratified sampling. Only a couple of the tools in the WKBIOPTIM suite take the past sampling design, including stratification, into account and it would be beneficial to implement the possibility to do so.

3 Sampling with and without replacement

To think about resampling with or without replacement, assume a dataset (S0) was collected from a larger population (P). If S0 were infinitely large, then every estimator (e.g. mean or length frequency distribution) from P would be known with perfect precision and no uncertainty. In general, for any size S0, the statistics about S0 can be calculated with a perfect precision, e.g. because the mean of S0 is known. However, because S0 is not infinitely large, we do not know the exact mean (and other estimators) of P. To calculate the uncertainty of statistics for P, it is necessary to resample with replacement (i.e. bootstrap) the S0 dataset, to represent the distribution of P that we don't know but is assumed to be similar to S0. Therefore, if it is necessary to make inference about P, then the resampling should be done with replacement. Alternatively, if the scope of inference is S0 (i.e. S0 is the extent over which the inferences are to apply), then we have all the information in that sample, S0. We can ask what might happen if we had collected a dataset smaller than S0 by looking at random subsets/subsamples of S0, i.e. resample without replacement. If we want to ask what would happen if we repeated the process of taking a sample from P that was smaller, then we would need to resample with replacement.

3.1 Literature review

The bootstrapping technique was first considered in a systematic manner by Efron (1979).

The essence of bootstrapping is the idea that, in absence of any other knowledge about a population, the distribution of values found in a random sample of size n from the population is the best guide to the distribution in the population (Manly, 1997).

The infinite population that consists of the n observed sample values, each with 1/n probability to be extracted, is used to model the unknown real population. The sampling is with replacement, which is the only difference in practice between bootstrapping and randomization in many applications.

When we sample with replacement, the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second. Mathematically, this means that the covariance between the two is zero.

In sampling without replacement, the two sample values aren't independent.

The Recycling Rate (RR) is an indicator of the percentage of re-use of the same sample when running the bootstrap; in the sampling with replacing, the RR is more affected.

In cases of a huge number of samples to be re-sampled the 2 approaches (with and without replacement) are expected to provide similar results, because the probability of extraction of each sample in the cases become very similar (1/n similar to 1/(n-1) when n is big). The test on a simulated population has been carried out to explore how the 2 approches work on the same artificial dataset, on which the real characteristics were known.

3.2 Case studies applying a subsampling with and without replacement

A case study applying SampleOptim to test the sampling with and without replacement, in order to compare the outputs of sampling optimisation output on real sample data, is presented in section 4.2.

A preliminary simulation study, using simulated data, was performed applying different subsampling designs (with and without replacement) to evaluate the accuracy of the estimates of mean and standard deviation for length-at-age. The results from this application are presented in section 4.3.

4 Case studies

4.1 Admissible Dissimilarity Value (ADV)[3] – The case study of Red Mullet in the Aegean Sea (GSA 22)

The ADV analysis was performed on Red Mullet in the Aegean Sea (GSA 22) for the year 2019 and for the 1st quarter. The data were collected as part of the Greek Data Collection Framework onboard sampling. For bottom – trawlers, biological data are collected through onboard sampling and the sampling scheme is based on fishing trips. Both landings and discards are recorded for catches and individual lengths are recorded for almost all species caught.

ADV is a tool described in Wischnewski et al. (2020) for comparing an LFD with sub-samples of itself in order to calculate the minimum subsample that holds the same properties as the original one.

4.1.1 Construction of the reference subsample

In order to minimize the number of samples needed to be taken and at the same time preserve the characteristics of the original LFD we need to construct a theoretical reference sub sample. This sub sample is constructed as follows:

First we need to choose the bandwidth of the LFD in a way that the new LFD will not lose the characteristics of the original one.

We also need to define a parameter γ which is a threshold up to which we can reduce the number of individuals in each length class, e.g. if $\gamma = 0.2$ we are allowed to reduce each length class up to 20% of each original number of individuals.

A proper value for the parameters ε and θ must also be chosen. The parameter ε depends on the species and it sets a threshold for the maximum distance that the modes (or antimodes) of the original sample from the modes (or antimodes) of the reference sample are allowed to have. θ is the maximum value of the amplitude ratio (Wischnewski et al. 2020).

For the case of Red Mullet in GSA 22 for 2019 and 4th quarter a bin width (Δ) = 2cm was selected. The ε is 0, $\gamma = 0.2$ and $\theta = 0.9$. The following figures (Figure 4.1.1.1. and 4.1.1.2) present the original LFD and the LFD with bin width = 2cm, respectively.

[3] SampleReferenceLevel R-tool

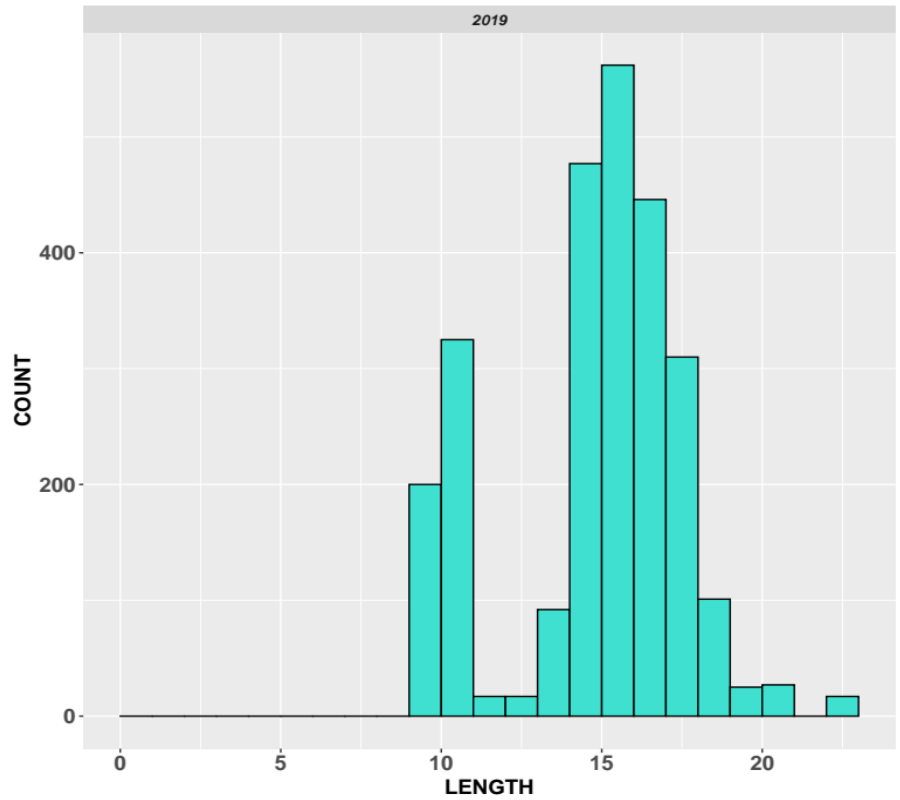


Figure 4.1.1.1. Histogram of length frequency distribution (LFD) of red mullet in GSA22 in 4th quarter, raised to the total catch with Δ = 1cm.

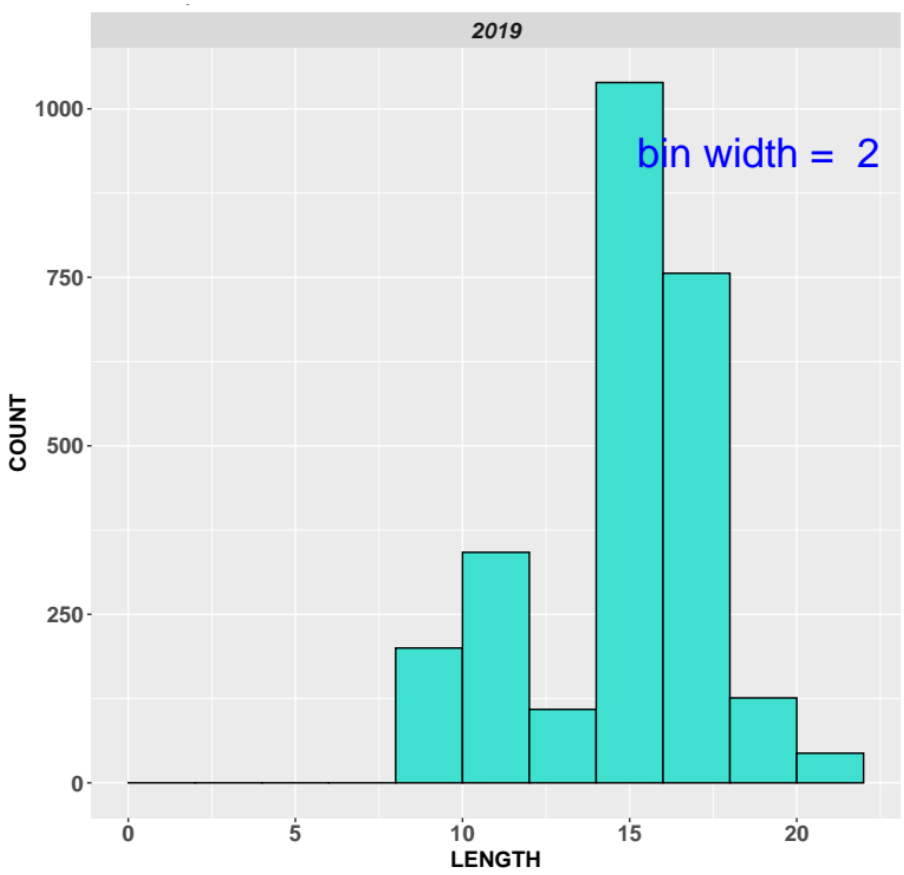


Figure 4.1.1.2. Histogram of length frequency distribution (LFD) of red mullet in GSA22 in 4th quarter, raised to the total catch with Δ = 2cm

Table 4.1.1. Results of *find_modes_antimodes* function for Δ= 1 and 2 cm.

	Δ = 1cm	**Δ = 2cm**
modes	10, 15, 20	10, 15, 20
antimodes	11, 19	11, 19
amplitudes	308, 545, 547, 2	308, 545, 537, 2
smoothed modes	10, 15, 20	10, 14
smoothed antimodes	11, 19, 21	12
robust modes	10, 15, 20	10,15
robust antimodes	11, 19	12
robust amplitudes	308, 545, 537, 2	308, 545

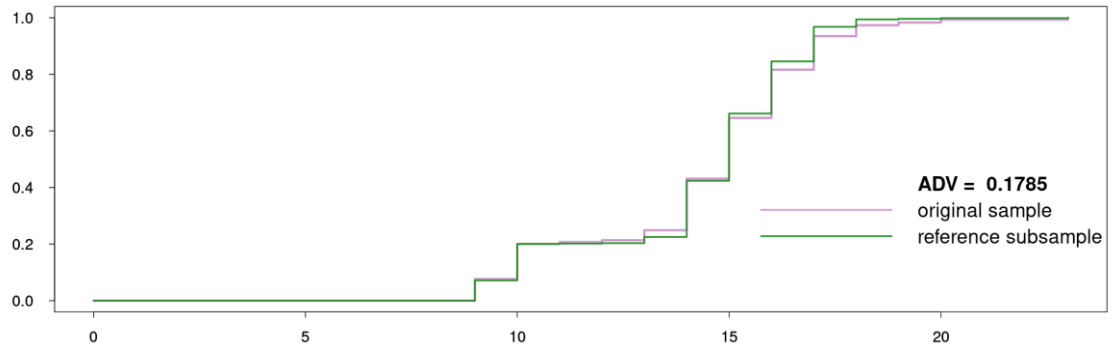


Figure 4.1.1.3. ADV of original sample and the reference subsample.

The ADV was calculated for the reference subsample as 0.1785 (Figure 4.1.1.3 and 4.1.1.4). The number of individuals in the original sample is 2616 while in the reference subsample 2180.

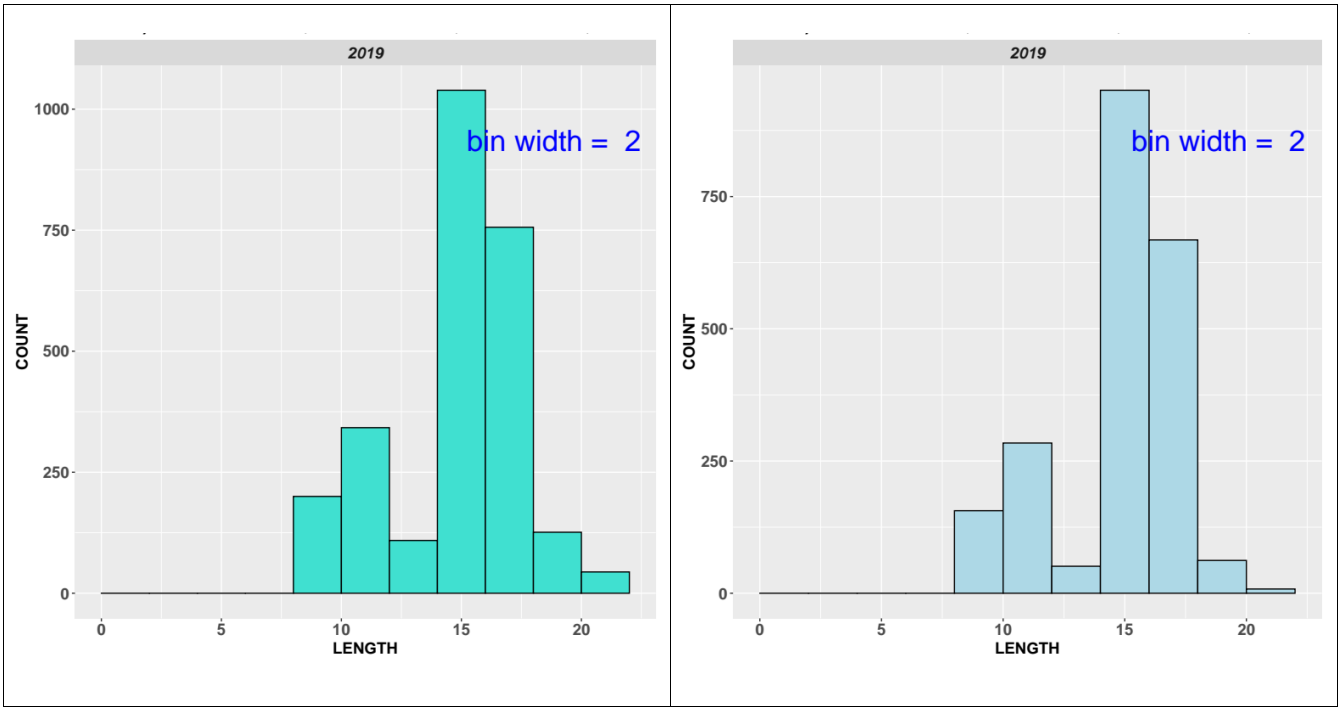


Figure 4.1.1.4. Original sample of red mullet in GSA22 in 4th quarter (left) and reference sub sample (right).

4.1.2 Reducing sampling effort by eliminating sampling units (fishing trips)

During the 4th quarter of 2019, six different trips were sampled. Namely trips: 9301, 9302, 11446, 11448, 11450 and 11456. We examine the effect on the sample by removing one, two and three trips.

In the case of removing one trip we will examine the effect of each trip in the sample size while for two and three we will simply remove randomly two and three trips from the original sample.

This results in eight different scenarios: $S_{1,j}$ where j are the fishing trips and S_2 and S_3 are the subsets of the original sample obtained by subtracting two and three trips respectively (Table 4.1.2).

Table 4.1.2. Results of different scenarios eliminating sampling units of the original sample.

Scenario	Eliminated Sampling Units	Distance D	Sample size
$S_{1,1}$	Trip 9301	12.2305	700
$S_{1,2}$	Trip 9302	0.0199	2607
$S_{1,3}$	Trip 11446	0.0391	2589
$S_{1,4}$	Trip 11448	0.0081	2607
$S_{1,5}$	Trip 11450	22.76224	1970
$S_{1,6}$	Trip 11456	0.0199	2607
S_2	Trips 9301 & 11456	12.1846	691
S_3	Trips 9302, 11450, 11456	22.81054	1952

In the case study of Red mullet in the Aegean Sea the samples are not equally distributed between trips. So the effort reduction by removing trips from the original sample has effect on the distance mainly based on the number of individuals measured on each trip. There are several trips that have almost no effect on the main characteristics of the LFD, but all of them contain very few individuals. The interesting part of this analysis is the effect of fishing trip 11450 on the distance. Although as a trip consists of fewer sampled individuals than trip 9301 the effect of removing it is much higher.

4.2 SampleOptim – The case study of Blue Whiting in ICES Division 27.6.a, a comparison of sampling with or without replacement

The aim of this case study was to investigate the effects of resampling strategy with replacement and without replacement on precision and bias of optimisation results, using real sample data.

In the case of sampling with replacement, all samples are independent as the selection of one sample does not affect the selection of another sample. When sampling without replacement, samples are not independent as the selection of one sample removes it from the sample pool available for subsequent sample selection.

The sampling optimisation approach SampleOptim was used, and quality indicators were generated to assess the optimal sample size for age-length keys.

Sampling optimisation was applied to age sample data of blue whiting (*Micromesistius poutassou*), stock whb.27.1-91214, from the 2021 spring season in ICES area 27.6.a.

4.2.1 Methods

The SampleOptim R-tool was run without replacement (scenario WOR) and with replacement (scenario WR) using the same input data. The input parameters for both scenarios are shown in Table 4.2.1.1. Only length classes that were sufficiently sampled (>30 age samples; 240-340 mm; Table 4.2.1.2) were selected for testing. The sample sizes tested were 2, 4, 6, 8, 10, 20 and 30 ages per length class; with sample sizes of 20 and 30 very close to the population size. No stratification was applied to the data. For each scenario, 1000 simulations were run to ensure stable output.

The output of the scenarios was compared by addressing two questions: i) Do the different scenarios result in significantly different output values? and ii) Do differences in the output of the two scenarios lead to different conclusions on sample size optimisation?

To compare the output values of mean lengths per age from the two scenarios, individual simulation output values were grouped by sample size and age (e.g. sample size 2 & age 2), and the results from each scenario was compared using an independent samples t-test. An alpha level of 0.01 was used for all statistical tests.

Bertalanffy parameter output values (Linf, K and t0) were not normally distributed, so the non-parametric independent 2-group Mann-Whitney U Test was used to compare the output results from the sample size groupings (e.g. sample size 2) of each scenario.

A visual assessment of the graphical output of the SampleOptim tool is used to make judgements on sample size optimisation. Therefore, the graphical outputs from the two scenarios were compared to assess for any differing conclusions of optimal sample sizes.

Table 4.2.1.1. Input parameters.

Variable name	*Mandatory*	*Variable.values*	*Definition*
species	y	WHB	CODE_FAO
AREA	y		
AGE_ONLY	y	TRUE	TRUE- only statistical analysis for age; FALSE - includes age and maturity data analysis
PORT	n	FALSE	Uses Port stratification for subsampling (TRUE); do not take into account the Port Stratification (FALSE)
distUniPorto	n	FALSE	Uniform distribution of subsamples by Port (TRUE); random distribution of subsamples by Port (FALSE)
TIME_STRATA	y	A	A - year; S- semester; T - quarter
SEX_RATIO	y	-1	0 - only males; 1 - only females; -1 no sex data
MIN_LC	y	240	minimum length class
MAX_LC	y	340	maximum length class
interval_LC	y	10	length class step
MIN_age	y	1	minimum age
MAX_age	y	10	maximum age
MIN_OTOL	y	2	minimum number of individuals by length class in the simulation setup
MAX_OTOL	y	10	maximum number of individuals by length class in the simulation setup
interval_OTOL	y	2	interval number of individuals by length class in the simulation setup
EXTRA_OTOL	n	20 30	extra otoliths that are not within MIN_OTOL and MAX_OTOL
Linf	y	45	Von Bertalanffy growth model parameter - Linf. Used as a starting value to adjust VBGM.
K	y	0.1	Von Bertalanffy growth model parameter - k. Used as a starting value to adjust VBGM.
t0	y	-3	Von Bertalanffy growth model parameter - t0. Used as a starting value to adjust VBGM.
year_start	y	2021	first year data subset to run simulations
year_end	y	2021	last year data subset to run simulations
stage_mature	y	2	define the maturity stages that correspond to mature stages (to allow to determine the proportion of immatures and matures)
n	y	1000	define the number of simulations (bootstrap runs)

Table 4.2.1.2. Selected input data for all simulations, corresponding to the input parameters (Table 4.2.1.1).

Length Class	240	250	260	270	280	290	300	310	320	330	340
N. Oto	35	56	65	75	76	79	80	73	63	40	31

4.2.2 Results and Discussion

T-test results comparing the mean length output, grouped by sample size and age, for scenarios WOR and WR showed that there were only two cases in which statistically significant differences ($p \geq 0.01$) in mean length when samples sizes were 10 or fewer per length class. However, for sample sizes of 20 and 30, differences in the mean length were statistically significantly between the WOR and WR scenarios for all ages ($p<0.01$), except age 3 for a sample size of 20 ($p>0.01$) (Table 4.2.2.1).

Comparison of the von Bertalanffy growth parameter Linf generated for the WOR and WR scenarios showed a statistically significant difference in values for samples sizes of 4, 8, 20 and 30 ($p< 0.01$). Statistically significant differences in K value between the WOR and WR scenarios was found for sample sizes of 4, 20 and 30 ($p< 0.01$), and no significant differences were found between the t0 values generated in each scenario (Table 4.2.2.2).

Visual inspection of the SampleOptim results is used to judge the point at which values stabilise, indicating that a sufficient sample size has been reached. For both the WOR and WR scenarios, the plotted results of the mean length at age stabilised at 8-10 otoliths (Figure 4.2.2.1). Similarly, the results of the von Bertalanffy growth parameter results showed that collecting about 8-10 otoliths would sufficiently stabilize the estimates of Linf, K and t0, for both the WOR and the WR scenarios (Figure 4.2.2.2).

The results of this case study show that applying the sample optimisation tool SampleOptim with and without replacement can have a significant effect on the calculation of both the mean length at age and the von Bertalanffy parameters Linf and K, particularly when the sample size is close to the maximum available in samples. This result is not surprising given that sampling without replacement places greater constraint on the pool available to be sampled as sample sizes get closer to the size of the population being sampled. As all samples are independent when sampling with replacement, there is no such constraint of sample sizes being near the sampled population size.

However, despite differences in output values at large sample sizes, the case study also showed that the resulting selection of optimal sample size was unaffected by the application of sampling with or without replacement.

Table 4.2.2.1. T-test results comparing mean length per sample size and age for scenarios WOR and WR for sample sizes. Age 10 is excluded from calculations because only one single otolith was available for this age.

Sample size	Age	T	df	P-value	Scenario WOR	Scenario WR
					Mean length (standard error)	Mean length (standard error)
2	2	0.7	1963.4	0.495	247.29 (±0.20)	247.10 (±0.19)
	3	-0.2	1980.3	0.806	266.44 (±0.35)	266.56 (±0.34)
	4	-0.3	1977.3	0.767	291.15 (±0.46)	291.34 (±0.48)
	5	-0.6	1990.8	0.537	304.81 (±0.36)	305.12 (±0.34)
	6	0.9	1931.9	0.382	303.76 (±0.44)	303.21 (±0.45)
	7	0.7	1839.8	0.481	312.85 (±0.49)	312.36 (±0.50)
	8	-1.2	1004.9	0.242	312.83 (±0.82)	314.19 (±0.82)
	9	1.7	502.8	0.087	319.88 (±0.93)	317.66 (±0.90)
4	2	-0.1	1992.7	0.914	247.86 (±0.13)	247.89 (±0.14)
	3	-0.1	1991.7	0.940	266.88 (±0.23)	266.91 (±0.22)
	4	0.6	1999.0	0.541	291.35 (±0.31)	291.08 (±0.31)
	5	-1.3	1998.1	0.211	304.55 (±0.23)	304.97 (±0.24)
	6	1.3	1992.7	0.195	304.19 (±0.30)	303.66 (±0.28)
	7	2.0	1982.6	0.050	313.35 (±0.34)	312.42 (±0.33)
	8	0.6	1520.8	0.530	314.25 (±0.59)	313.71 (±0.61)
	9	0.5	857.4	0.640	318.25 (±0.65)	317.82 (±0.66)
6	2	0.3	1993.7	0.792	247.89 (±0.11)	247.85 (±0.11)
	3	-0.9	1993.3	0.370	266.80 (±0.17)	267.02 (±0.18)
	4	1.0	1993.6	0.338	291.56 (±0.23)	291.24 (±0.24)
	5	0.1	1998.6	0.922	304.37 (±0.18)	304.35 (±0.18)
	6	0.6	1999.9	0.573	304.07 (±0.22)	303.89 (±0.22)
	7	1.6	1992.2	0.118	312.89 (±0.27)	312.31 (±0.26)
	8	1.1	1750.1	0.270	313.96 (±0.49)	313.18 (±0.51)
	9	0.3	1165.9	0.732	318.59 (±0.53)	318.33 (±0.55)
8	2	-1.2	1993.7	0.221	247.87 (±0.09)	248.03 (±0.09)
	3	-0.4	1999.9	0.720	267.14 (±0.15)	267.22 (±0.15)
	4	0.7	1996.9	0.497	291.54 (±0.20)	291.34 (±0.21)
	5	2.8	1999.8	0.005	304.73 (±0.16)	304.10 (±0.16)
	6	1.5	1986.6	0.137	303.78 (±0.18)	303.38 (±0.20)
	7	1.4	1993.2	0.154	312.85 (±0.21)	312.41 (±0.23)
	8	1.2	1868.8	0.229	314.20 (±0.42)	313.47 (±0.44)
	9	1.6	1393.6	0.107	318.90 (±0.47)	317.80 (±0.49)
10	2	-2.6	1975.1	0.008	247.81 (±0.08)	248.12 (±0.09)
	3	-1.3	1999.7	0.182	266.96 (±0.13)	267.20 (±0.13)
	4	1.3	1984.4	0.191	291.69 (±0.16)	291.37 (±0.18)
	5	1.4	1991.2	0.158	304.60 (±0.13)	304.32 (±0.14)

Sample size	Age	T	df	P-value	Scenario WOR Mean length (standard error)	Scenario WR Mean length (standard error)
	6	2.4	1996.3	0.016	304.00 (±0.16)	303.46 (±0.16)
	7	1.6	1994	0.108	312.34 (±0.18)	311.91 (±0.20)
	8	1.5	1931.7	0.133	313.69 (±0.37)	312.88 (±0.40)
	9	1.7	1499.5	0.096	319.36 (±0.43)	318.31 (±0.46)
20	2	-9.6	1894.9	<0.001	247.91 (±0.05)	248.68 (±0.06)
	3	-2.2	1977.5	0.026	267.16 (±0.08)	267.41 (±0.09)
	4	3.3	1977.4	0.001	291.78 (±0.10)	291.29 (±0.11)
	5	8.4	1958.9	<0.001	304.58 (±0.08)	303.57 (±0.09)
	6	7.7	1982.5	<0.001	304.07 (±0.10)	302.98 (±0.10)
	7	8.4	1944.9	<0.001	312.38 (±0.11)	310.96 (±0.13)
	8	4.2	1920.8	<0.001	313.97 (±0.21)	312.58 (±0.26)
	9	4.6	1880	<0.001	320.06 (±0.29)	318.05 (±0.33)
30	2	-13.8	1766.1	<0.001	248.07 (±0.03)	248.93 (±0.05)
	3	-6.3	1798.5	<0.001	267.16 (±0.05)	267.67 (±0.07)
	4	7.9	1776.9	<0.001	291.76 (±0.06)	290.96 (±0.08)
	5	16.6	1875.5	<0.001	304.56 (±0.06)	303.06 (±0.07)
	6	11.3	1856.1	<0.001	303.95 (±0.06)	302.75 (±0.09)
	7	16.2	1798	<0.001	312.52 (±0.07)	310.57 (±0.10)
	8	8.3	1725.4	<0.001	313.80 (±0.13)	311.89 (±0.19)
	9	7	1722.2	<0.001	319.95 (±0.17)	317.85 (±0.25)

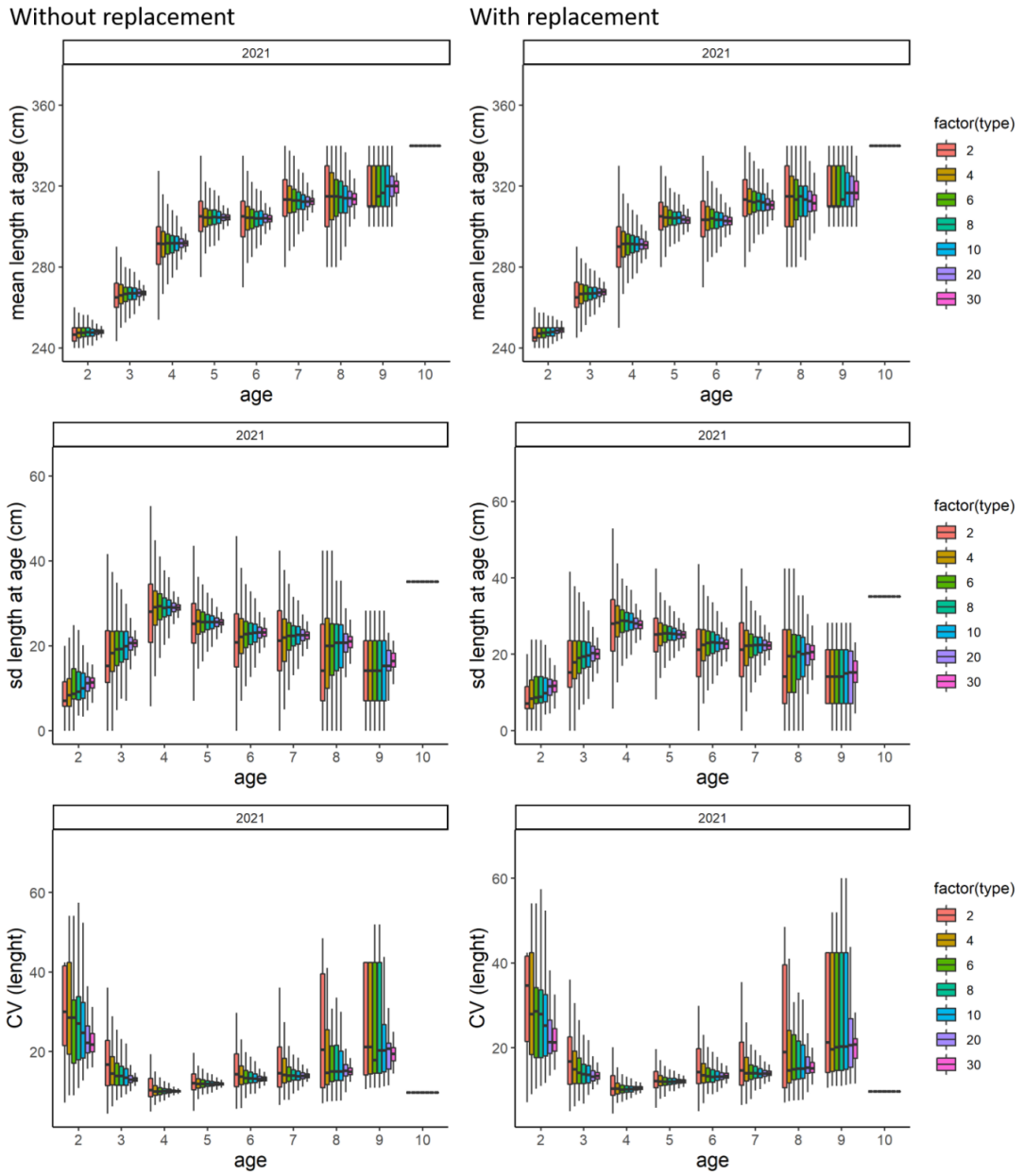


Figure 4.2.2.1. Comparison of the variability of mean length at age, standard deviation (sd) at age and coefficient of variation (CV) of sampling 2,4,6,8,10,20 and 30 otoliths per length interval, without replacement (WOR scenario; figures on the left) and with replacement (WR scenario; figures on the right). Note: a single sample contributed to age 10.

Table 4.2.2.2. U-test results comparing von Bertalanffy growth parameters Linf, K and t0 for scenarios WOR and WR at sample sizes 2, 4, 6, 8, 10, 20 and 30.

Sample size	Linf		K		t0	
	W	P-value	W	P-value	W	P-value
2	4528	0.080	3365	0.100	3335	0.083
4	6597	0.007	4267	0.008	4350	0.014
6	8106	0.149	6466	0.117	6642	0.213
8	12001	0.001	8391	0.047	8763	0.153
10	10945	0.374	10090	0.747	10468	0.832
20	17596	<0.001	9797	0.002	11577	0.369
30	7524	<0.001	2688	<0.001	3710	0.054

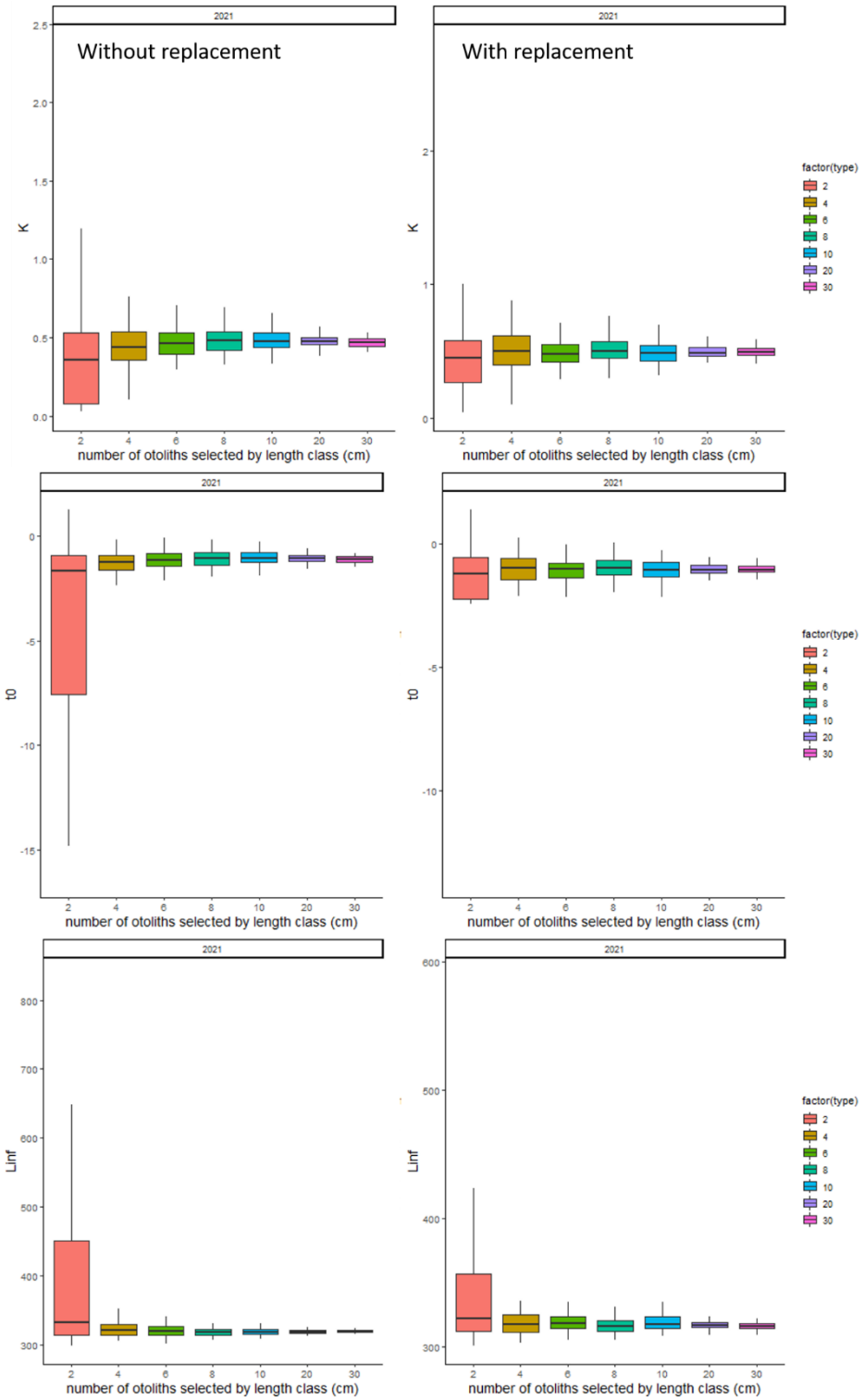


Figure 4.2.2.2. Comparison of the variability of von Bertalanffy coefficients of sampling 2,4,6,8,10,20 and 30 otoliths per length interval, without replacement (figures on the left) and with replacement (figures on the right).

4.3 Simulated population - applied to two different sub-sampling designs (with and without replacement)

A preliminary study using a simulated population has been performed to evaluate the accuracy on the estimates of mean and standard deviation for length-at-age by applying different subsampling designs, with and without replacement.

4.3.1 Methods

First, a population with a total of 10,000 fish per age, for ages 0 to 10, was simulated. Each fish was given a different value of Linf from a lognormal distribution with CV 0.05. This represents the population available to be caught. Then, length-at-age was calculated deterministically for each fish based on the von Bertalanffy growth model (Figure 4.3.1):

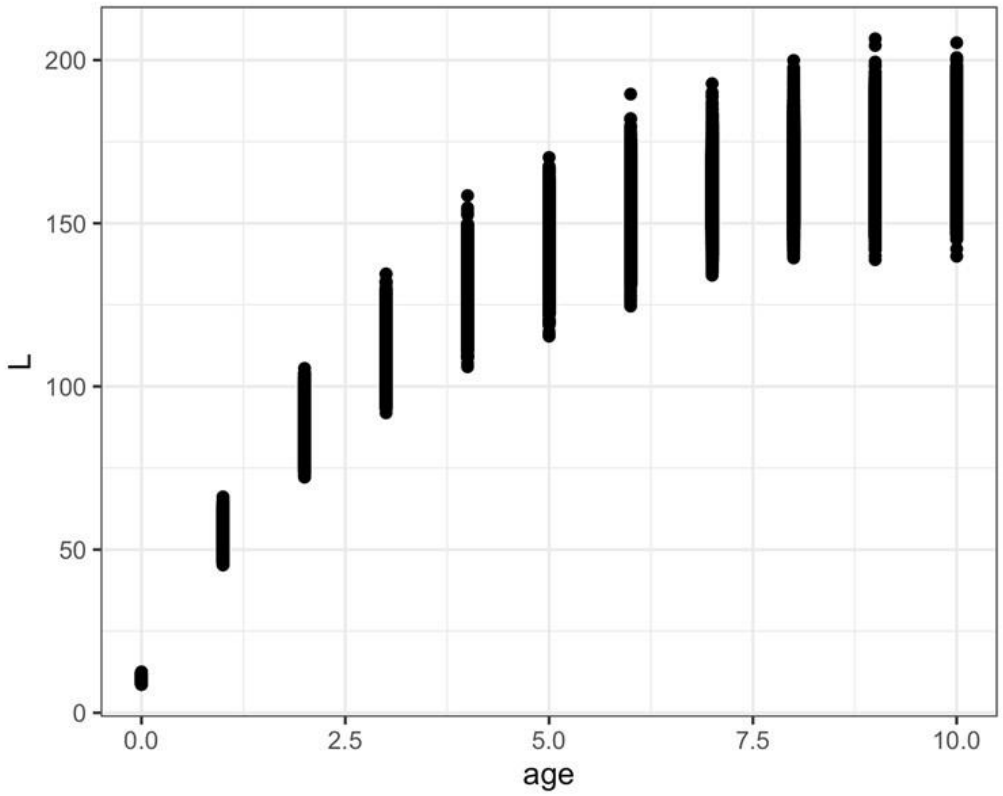


Figure 4.3.1. Simulated lengths at age. The population contains 10,000 individuals per age and individuals vary in their Linf parameter with a CV = 0.05. The other parameters of the von Bertalanffy growth model were t0=-0.2, k = 0.3, and mean Linf = 180.

A catch (S0, i.e. haul or landing) of size 50, 1000, or 10,000 fish were randomly sampled from the population described above. Then the catch was divided into 10 cm length classes. From each length class, nsamp (2, 5, 10, or 50) fish were randomly sampled with or without replacement, unless there were fewer than nsamp fish available and then only the number available were sampled. Based on these samples, the mean and standard deviation of the length-at-age were calculated. The procedure was repeated 1000 times for each sampling design (with and without replacement), i.e. 1000 replicate simulations per combination of S0 size and nsamp. Replicate

simulations with the same S0 size used the same fish, i.e. the catch did not change, only the samples taken from the catch changed.

4.3.2 Results

All sampling designs produced fairly accurate estimates of mean length-at-age (Figure 4.3.2.1).

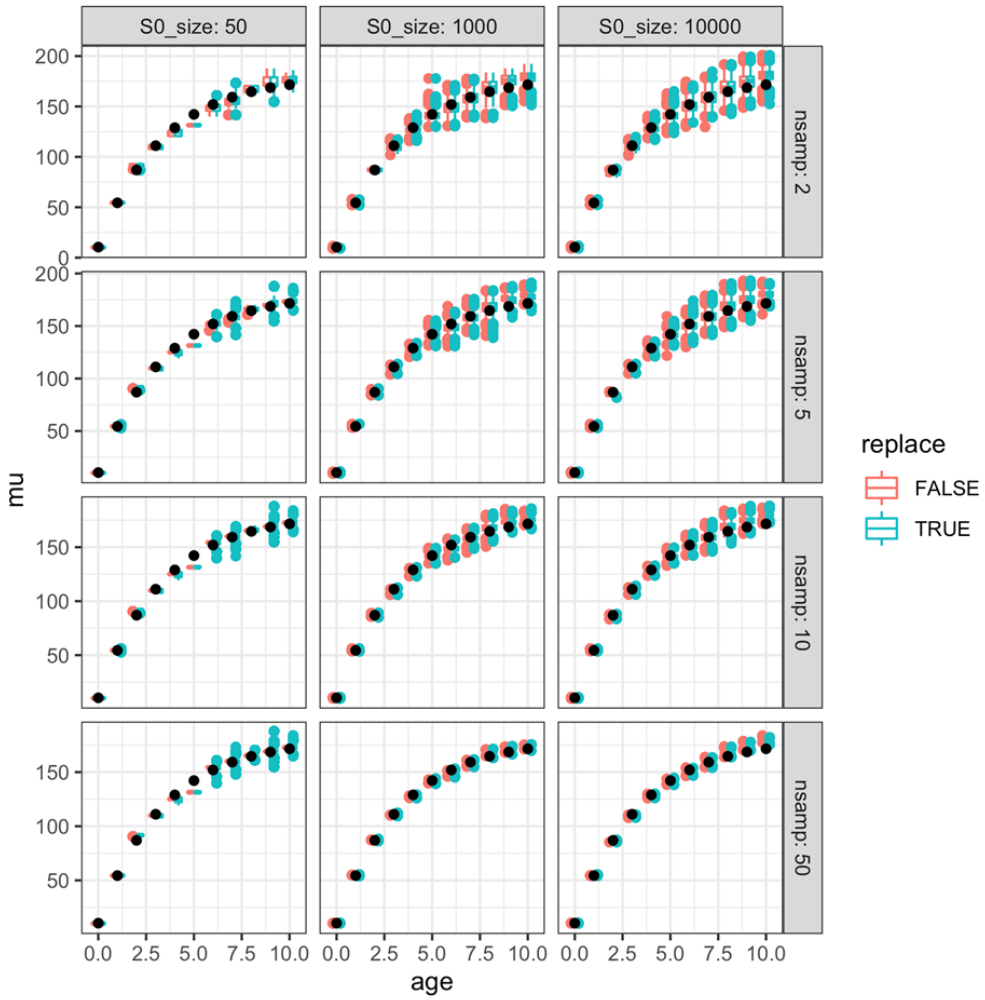


Figure 4.3.2.1. Mean length-at-age (mu). Black dots represent the true value based on the true growth model underlying the simulated data. S0_size varies by column and is the size of the catch or haul that was being sampled. Rows vary nsamp which is the number of samples per length class. Red and green boxplots show the values of 1000 replicates for each sampling design (without and with replacement).

Estimates of the standard deviation (sd) of length-at-age were mixed and sometimes biased (Figure 4.3.2.2). With a catch size of 50, there were too few samples to accurately estimate sd. With a catch size of 10,000, and nsamp equal to 10 or 50, estimates were inexplicably biased upward for higher age classes; more investigation is needed to confirm and explain this fact. With nsamp = 2 or 5, some of the replicates had too few samples in some age classes to be able to estimate a standard deviation for those age classes.

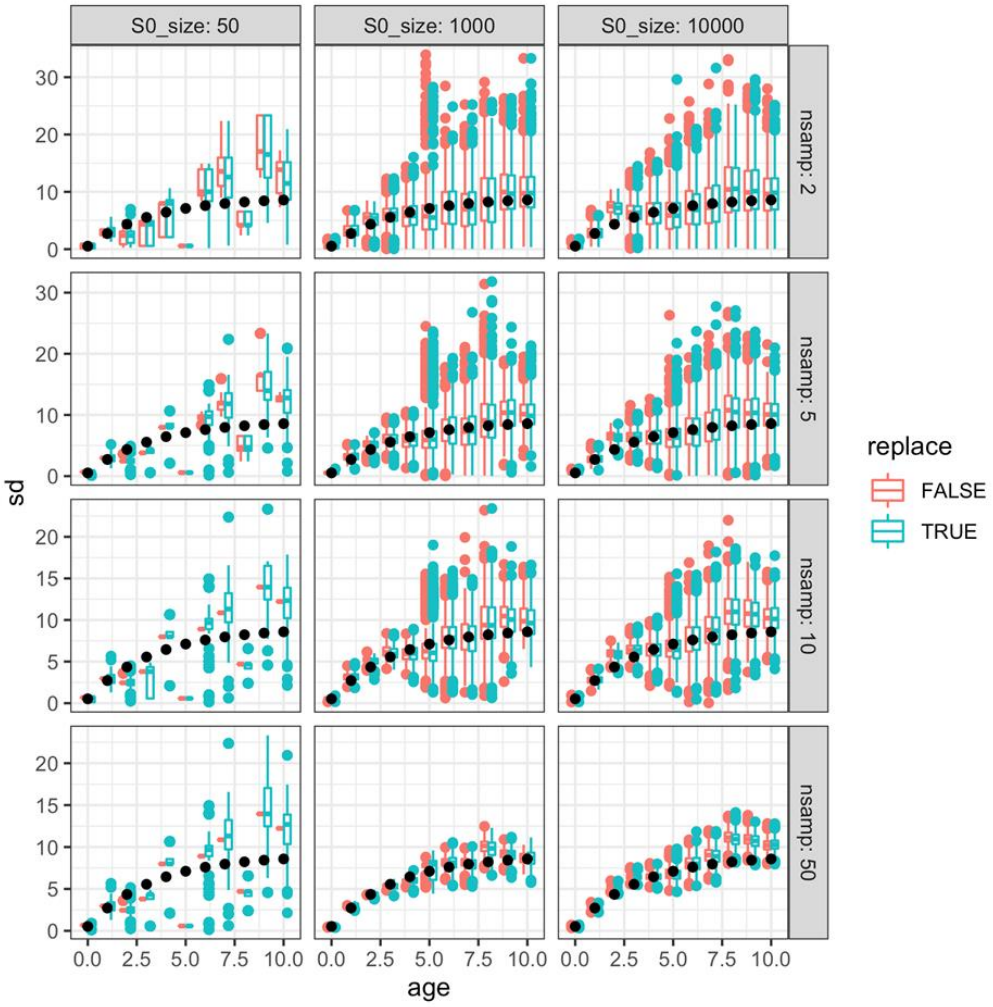


Figure 4.3.2.2. Standard deviation (sd) of length-at-age. Black dots represent the true value based on the true growth model underlying the simulated data. S0_size varies by column and is the size of the catch or haul that was being sampled. Rows vary nsamp which is the number of samples per length class. Red and green boxplots show the values of 1000 replicates for each sampling design (without and with replacement).

4.3.3 Discussion

On the current study, the same number per age group were simulated but, in future work, it potential consequences and an alternative approach should be evaluated.

More investigation is needed to make sure that these results are valid and to understand the bias of standard deviation estimates. Future work could also investigate age-length keys derived from this type of simulated data. The same analysis shown here for mean length-at-age should hold for mean weight-at-age, but this could be discussed at a future WKBIOPTIM meeting.

4.4 BioSim Tool – The case study of *Mullus barbatus* in GSA 22

4.4.1 Methods

BioSim Tool has been applied to the sampling data of *Mullus barbatus* in GSA 22 (Aegean Sea) on sex and maturity by length class. Sampling data from 2014 to 2020 have been utilized.

The focus of the analysis is on the sampling optimization of sex and maturity data; BioSim Tool was used for this analysis to have an idea of the impact on sampling precision by changing the number of individuals for which sex and maturity are collected by trip. The impact of different scenarios, from 10 to 100 sampled individuals by trip, is evaluated also on the sex ratio and maturity by length, compared with the original one in the data used as reference (population).

The number of simulations considered is 50 and the samples considered representative of the population has been set to have at least 20 individuals.

4.4.2 Results and discussion

The results showed that, in the hypothesis of an annual sampling, the precision in length composition is comparable to the one of the original sample already with 60 individuals measured by trip (Figure 4.4.2.1).

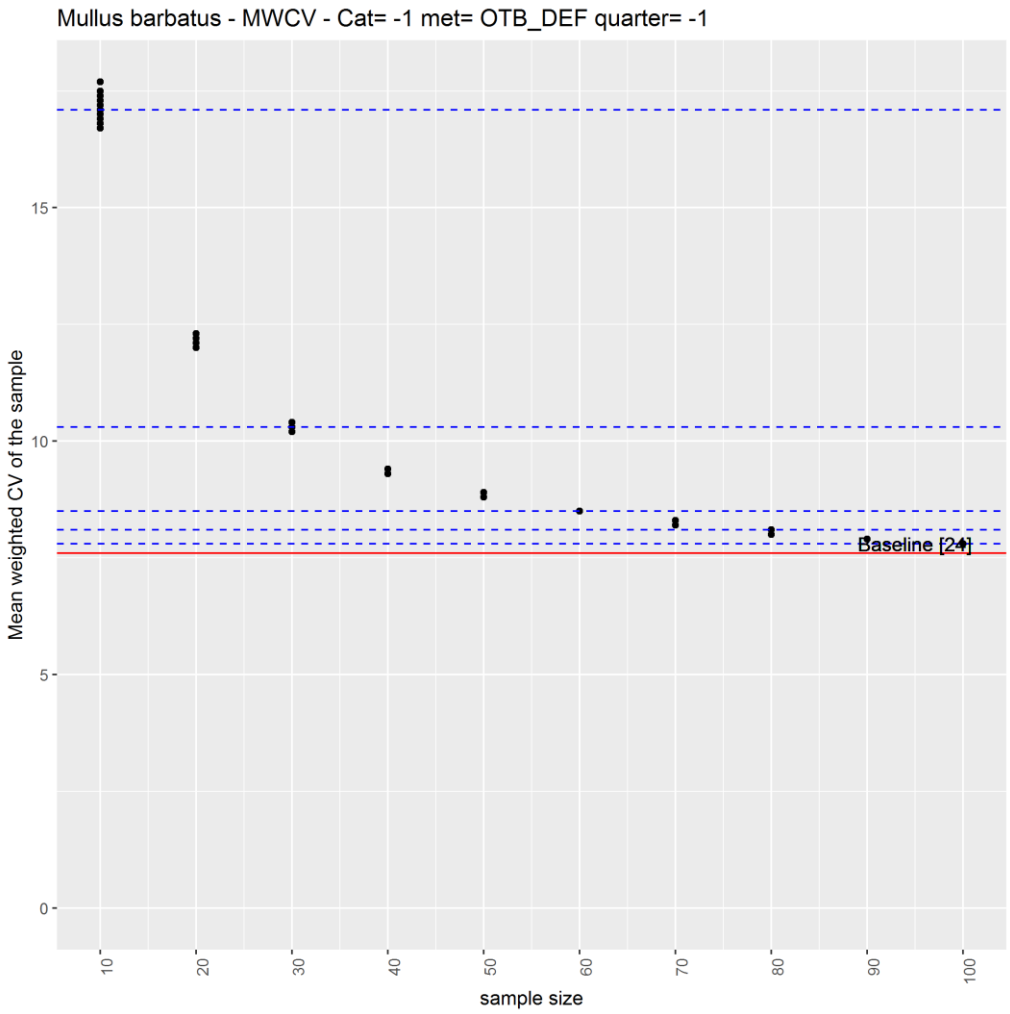


Fig 4.4.2.1 – CV of the length distribution in the hypothesis of 10, 20, …100 individuals sampled per trip.

On the other hand, the results of the sex showed that sampling about the same number (55-60 individuals) for the sex would produce a precision in the sex ratio very similar to the one of the original sample (Figure 4.4.2.2). Similar results are shown by the sex ratio at length (Figure 4.4.2.3).

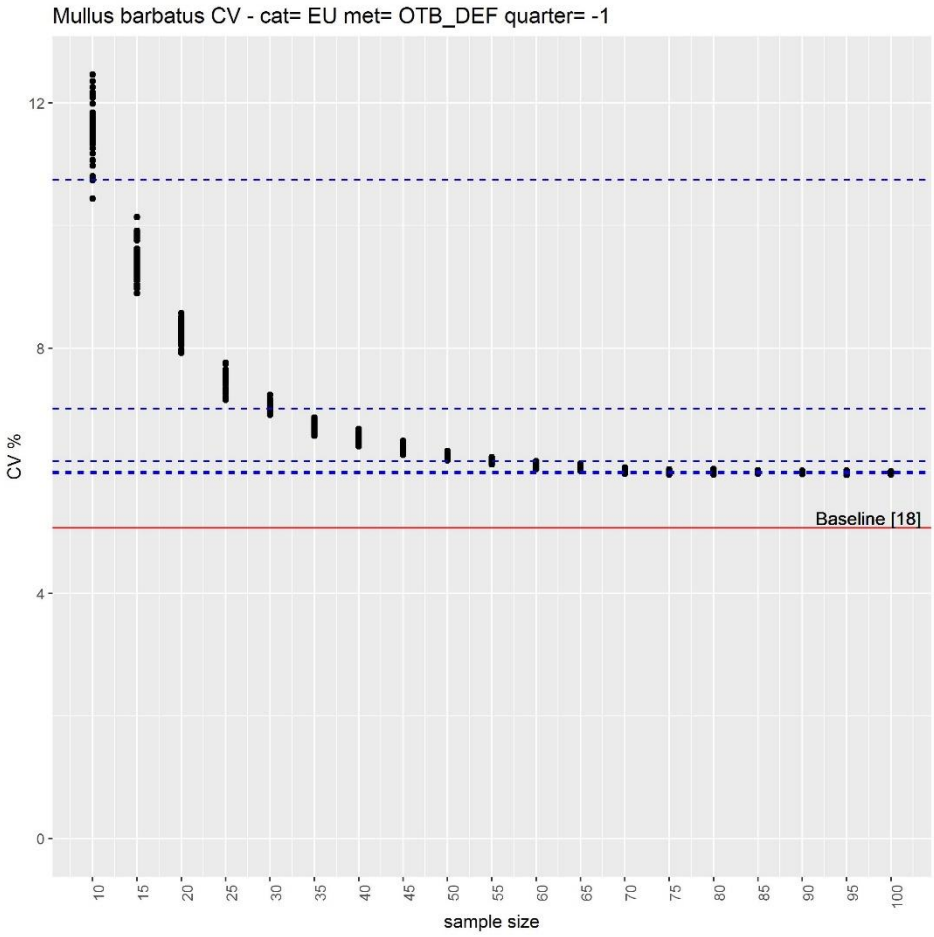


Figure 4.4.2.2 – CV of the sex ratio by length class in the hypothesis of 10, 20, …100 individuals sampled per trip for the sex.

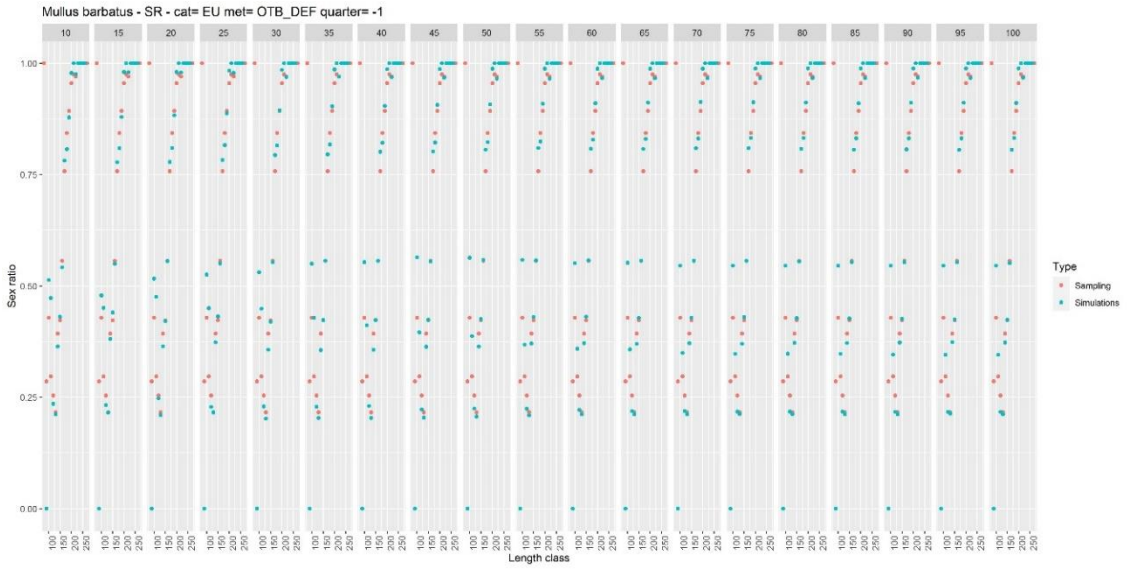


Figure 4.4.2.3 –Sex ratio by length class in the hypothesis of 10, 20, …100 individuals sampled per trip for the sex.

For length and sex variables the results of BioSim tool showed that a reduction of the number of individuals sampled for the length and for the sex would return a comparable precision and sex ratio estimate by length class.

5 Next steps

The next steps of the WKBIOPTIM's work will include:

[a] Continue working on the development and testing of the R-tools developed with the aim of providing support on fish sampling optimisation at national/stock/regional level;

[b] Evaluate the use of the R-tools for different sampling designs and where under-sampling may be occurring (e. g. small scale fisheries);

[c] Evaluate the compatibility between the different WKBIOPTIM R-tools, with the use of standard data formats and sources;

[d] Continue the R-package development;

[e] Working on the adaptation of the R-tools to accommodate the sampling schemes from the different hierarchies from the RDBES.

6 References

Efron, B. 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7, 1-26.

ICES. 2017. Report of the Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM), 20-22 June, Lisbon, Portugal. ICES CM 2017/SSGIEOM:32. 150pp.

ICES. 2019a. Report of the Workshop on Optimization of Biological Sampling (WKBIOPTIM 2). WKBIOPTIM 2 Report 2018 29–31 May 2018. Nantes, France. ICES CM 2018/EOSG:23. 172 pp.

ICES. 2019b. Workshop on Optimization of Biological Sampling (WKBIOPTIM 3). ICES Scientific Reports. 1:78. 219 pp. http://doi.org/10.17895/ices.pub.5647

ICES. 2022. Second Workshop on Estimation with the RDBES data model (WKRDB-EST2; outputs from 2020 meeting). ICES Scientific Reports. 3:15. 128 pp. https://doi.org/10.17895/ices.pub.7915

Manly, B.F.J. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall, London. 1997. 330p.

Wischnewski, J., Bernreuther, M., Kempf, A. 2020. Admissible dissimilarity value (ADV) as a measure of subsampling reliability: case study North Sea cod (*Gadus morhua*). Environmental Monitoring and Assessment 192, 756 (2020). https://doi.org/10.1007/s10661-020-08668-6

Annex 1: List of participants

Name	Institute	Country (of institute)	Email
Ana Cláudia Fernandes	IPMA - Portuguese Institute for Ocean and Atmosphere	Portugal	acfernandes@ipma.pt
Danai Mantopoulou Palouka	HCMR - Hellenic Centre for Marine Research	Greece	danaim@hcmr.gr
Gwladys Lambert (chair)	Centre for Environment Fisheries and Aq-uaculture Science (Cefas)	UK	gwladys.lambert@cefas.co.uk
Ioannis Ntokos	HCMR - Hellenic Centre for Marine Research	Greece	gdokos@hcmr.gr
Isabella Bitetto (chair)	COISPA Tecnologia & Ricerca, Stazione Sperimentale per lo Studio delle Risorse del Mare	Italy	bitetto@coispa.it
Jessica Craig	Marine Laboratory Scotland	UK	Jessica.Craig@gov.scot
Julia Wischnewski	Thünen Institute of Sea Fisheries	Germany	julia.wischnewski@thuenen.de
Kirsten Birch Håkansson	DTU Aqua - National Institute of Aquatic Resources	Denmark	kih@aqua.dtu.dk
Mollie Elizabeth Brooks	DTU Aqua - National Institute of Aquatic Resources	Denmark	molbr@aqua.dtu.dk
Nicholas Carey	Marine Laboratory Scotland	UK	Nicholas.Carey@gov.scot
Patrícia Gonçalves (chair)	IPMA - Portuguese Institute for Ocean and Atmosphere	Portugal	patricia@ipma.pt

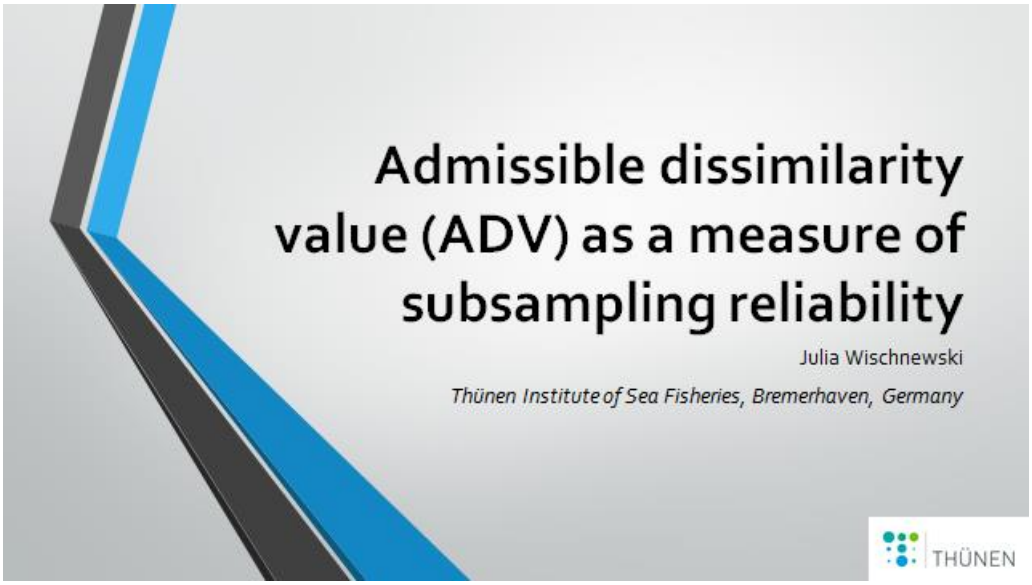
Annex 2: Agenda

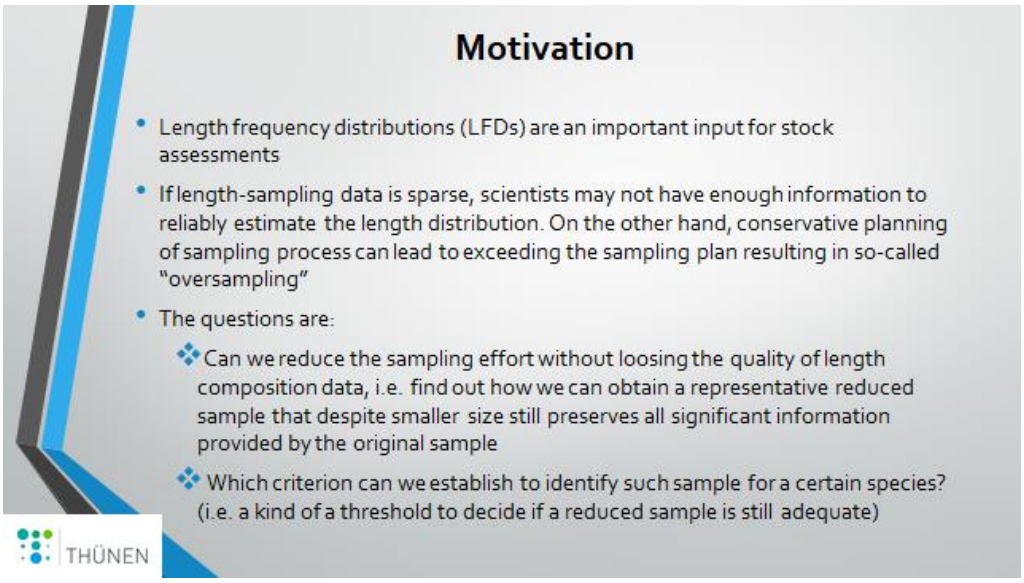
Monday 15th November		
10:00-10:15	Welcome and logistics	Chairs
10:15-10:40	Presentation of ToR's and adoption of agenda	Chairs
10:40-11:00	Presentation on the available WKBIOPTIM tools and on their possible integration	Patrícia Gonçalves Isabella Bitetto
11:00-11:30	Discussion on tools integration	Plenary
11:30 –Coffee break		
11:50-12:20	Presentation of RDBES	Henrik Kjems-Nielsen
12:20-13:00	Presentationabout an R package construction (roxygen)	Gwladys Lambert
13:00 - Lunch break		
14:00-15:00	Discussion on steps to convert the input data from RDB format to RDBES format	Plenary
15:00-15:30	Subgroups and work organization	Plenary
15:30 - Coffee break		
15:40 – 16:00	Wrap up of the day	Plenary
Tuesday 16thNovember		
10:00-10:20	New developments SampleOptim	Patrícia Gonçalves
10:20-10:30	New developments on STREAM optimization tools within STREAMline project	Isabella Bitetto
10:30-11:00	Admissible dissimilarity value (ADV) as a measure of subsampling reliability	Julia Wischnewski
11:00-11:30	Presentation on the quality indicators	Gwladys Lambert
11:30 –Coffee break		
11:50- 12:10	More on quality indicators - Coefficient of variation	Mollie Brooks
12:10-13:00	Discuss tool integration and documentation, define subgroups and roles.	Plenary
13:00 - Lunch break		
14:00 – 15:30	Subgroup work	Subgroups
15:30 - Coffee break		
15:40 – 16:00	Wrap up of the day	Plenary

Wednesday 17th November		
10:00-11:00	Sampling with or without replacement	Patrícia Gonçalves Isabella Bitetto
11:00-11:30	Definition of subgroups to test the impact on with or without replacement sampling in the R tools	Plenary
11:30 –Coffee break		
11:50-13:00	Subgroup work	Subgroups
13:00 - Lunch break		
14:00 – 15:30	Subgroup work	Subgroups
15:30 - Coffee break		
15:40 – 16:00	Wrap up of the day	Plenary
Thursday 18th November		
10:00-10:30	Link with WGCATCH	Liz Clarke
10:30-10:50	Update on R package development	Plenary
10:50-11:30	Subgroup work	Subgroups
11:30 - Coffee break		
11:50-13:00	Subgroup work	Subgroups
13:00 - Lunch break		
14:00 – 15:00	Subgroup work	Subgroups
15:00 – 15:30	Presentation and discussion of output/work done by sub-groups	Plenary
15:30 - Coffee break		
15:40 – 16:30	Report preparation	Plenary
Friday 19st November		
10:00-11:30	Presentation and discussion of output/work done by subgroups	Plenary
11:30 Coffee break		
11:30 – 12:00	Report preparation and discussion of future work	Plenary
12:00 – 13:00	Wrap up of the meeting	Plenary

Annex 3: Presentations

1. SampleReferenceLevel (ADV) (Section 1.4.1)





Criterion of the sample size adequateness

- Entire shape of LFDs is more important than descriptive statistics like the mean length, variance etc. So, as criterion that the reduced sample is still reliable (i.e. distributionally similar to original sample) we use a function based on a difference between LFDs (or CDFs) structure of reduced sample S_r (or subsample) and original sample S_o (or benchmark sample)
- This function is constructed on the base of metric distance between CDFs
- The minimally achievable subsample ("the worst case from all admissible/acceptable ones") S_r^{min}, which still satisfies the criterion, defines a threshold called Admissible Dissimilarity Value (ADV). It is basically a distance between CDFs of S_o and S_r^{min}
- All subsamples delivering distances greater than ADV are labelled as unreliable/inadequate

THÜNEN

Similarity interpretation

Distributions (of original sample and subsample) can be considered as similar if:

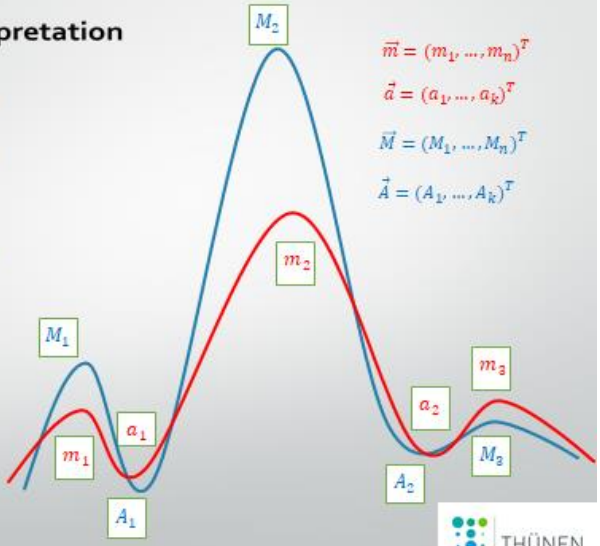
1) They have the same number of modes and antimodes, i.e. $\dim(\vec{m})=\dim(\vec{M})$ and $\dim(\vec{a})=\dim(\vec{A})$
2) For each corresponding pair m_i, M_i and a_j, A_j: $|m_i - M_i| \leq \varepsilon$ and $|a_j - A_j| \leq \varepsilon$, where $\varepsilon = \varepsilon(\text{species})$
3) amplitudes ratio $\frac{|g(m_i)-g(a_j)|}{|f(M_i)-f(A_j)|} \geq \theta$, where $f(\cdot)$, $g(\cdot)$ are the values of the original and reduced sampled density functions (LFD) at a point, $0<\theta \leq 1$

$\vec{m} = (m_1, \ldots, m_n)^T$

$\vec{a} = (a_1, \ldots, a_k)^T$

$\vec{M} = (M_1, \ldots, M_n)^T$

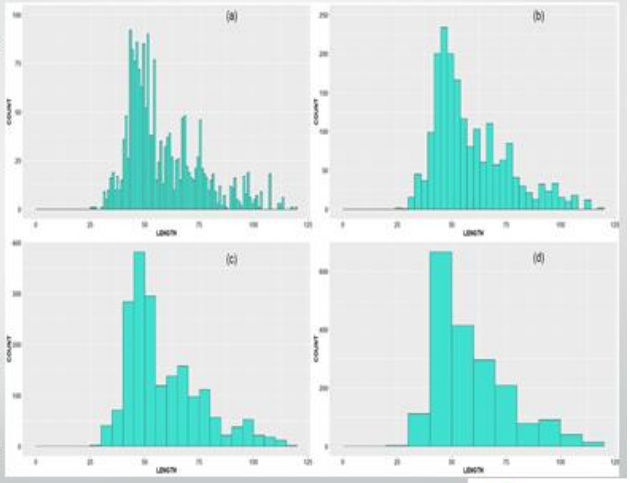
$\vec{A} = (A_1, \ldots, A_k)^T$



THÜNEN

Similarity interpretation

Note that we are focused only on robust modes/antimodes (Definition 1 in the manuscript) – i.e. modes/antimodes, which continue to be present despite length class smoothing and don't „contaminate" the distributional shape



LFD (raised to whole catch) of North Sea cod, obtained by the German commercial observers in the 3rd quarter of 2018, in ICES area 27.4: original sample with different bandwidth: (a) bandwidth = 1 cm; (b) bandwidth = 3 cm; (c) bandwidth = 5 cm; (d) bandwidth = 10 cm.

THÜNEN

Formal dissimilarity criterion

$$D(S_o, S_r) = L_1(F, G) + c_1 \cdot \mathbb{1}\{\dim(\vec{v}) \neq \dim(\vec{V})\} +$$

$$c_2 \cdot \sum_{i=1}^{dim(\vec{v})} \max(0, |v_i - V_i| - \varepsilon) \cdot \mathbb{1}\{\dim(\vec{v}) = \dim(\vec{V})\} +$$

$$c_3 \cdot \sum_{i=2}^{dim(\vec{v})} \max\left(0, \theta - \frac{|g(v_i) - g(v_{i-1}))|}{|f(v_i) - f(v_{i-1}))|}\right) \cdot \mathbb{1}\{\dim(\vec{v}) = \dim(\vec{V})\},$$

where F and G are CDFs of S_o and S_r,

$L_1(F, G)$ is the L_1-distance between them,

$\vec{V} = sort(\vec{M}, \vec{A})$ and $\vec{v} = sort(\vec{m}, \vec{a})$ be the sorted increasing sequences of the robust modes/antimodes of S_o and S_r, respectively,

f and g are density functions of S_o and S_r,

c_1, c_2 and c_3 are some constants

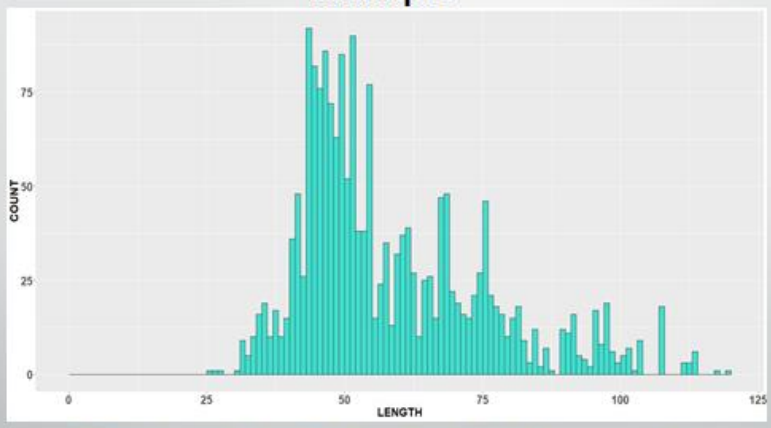
THÜNEN

Formal dissimilarity criterion

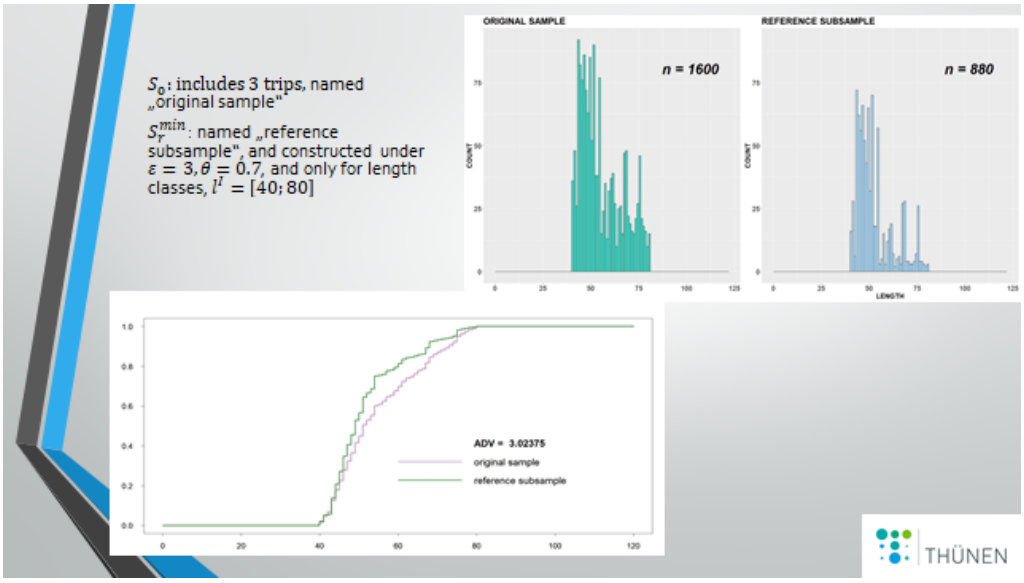
- Obviously, for $S_r \equiv S_o$ we obtain the lower bound $D = 0$
- The upper bound is the ADV and provided by the S_r^{min}
- So, $D(S_o, S_r) \in [0; \text{ADV}]$ for all acceptable S_r
- We should make clear, that the subsample is a purely theoretical and formally constructed on the base of the conditions (1)-(3). We use it only to conclude about the reliability of the derived "real-world" subsamples with respect to the original sample
- The subsample S_r^{min} and ADV is produced by the iterative subsampling procedure
- Note, that it is not always necessary, to be focused on information delivered by all length classes $l = \{l_j\}, j = 1,2, \ldots L$. One can choose the most important lengths classes $l^I = \{l_{j \in I}\}, I \subseteq \{1,2, \ldots, L\}$ where the most significant information is provided while less informative ones are ignored (e.g. very small or very large fish)

THÜNEN

Example



LFD (raised to wholecatch) of North Sea cod, obtained by the German commercial observers in the 3rd quarter 2018, in the ICES area 27.4: original sample.

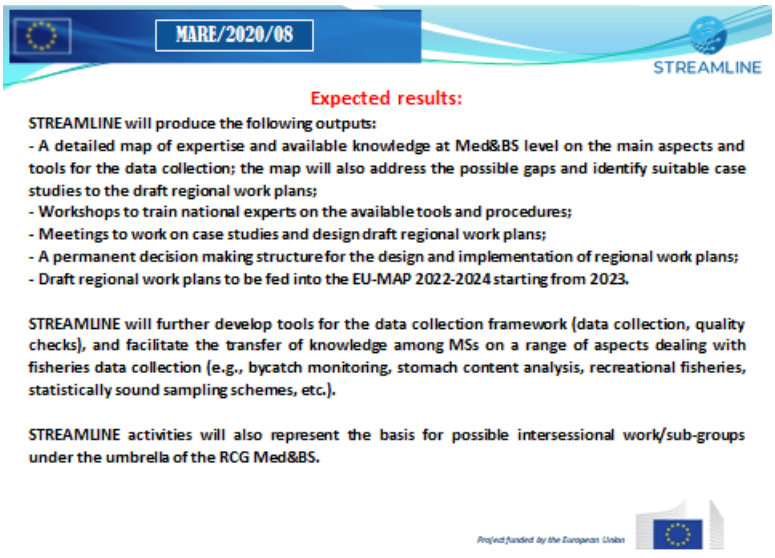


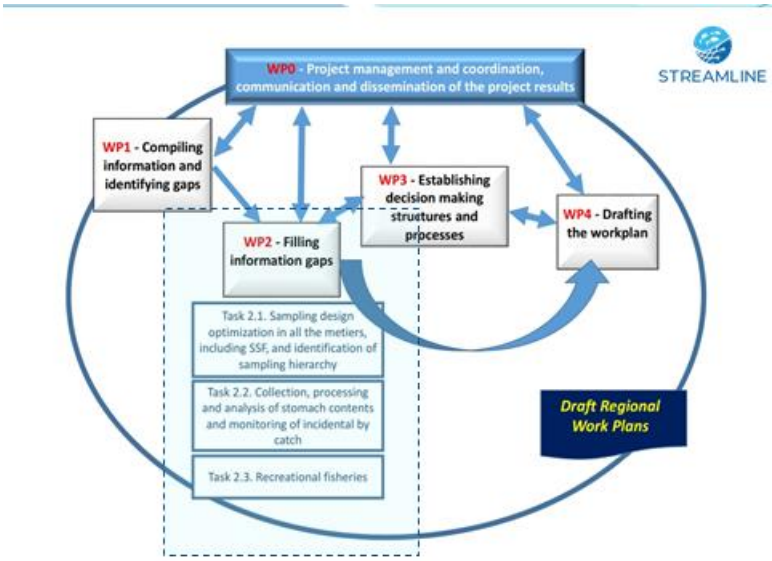
Example: Sampling units elimination (effort reducing), based on ADV

S_r	Eliminated sampling units	Distance D	Sample size, all length classes / length classes $l^I = [40;80]$	Mean length / standard error of mean length in length classes $l^I = [40;80]$
S_0	None	0	1924 / 1600	55.99 / 1.77
Trips elimination				
S_1	Trip 1	0.4744	1657 / 1428	54.14 / 1.08
S_2	Trip 2	12.8497	720 / 525	57.65 / 1.92
S_3	Trip 3	0.6616	1471 / 1247	56.31 / 2.87
S_4	Trip 1 + Trip 3	11.3927	1204 / 1075	53.63 / 1.54
Hauls elimination				
S_{10}	ω Trip 1 ω night time hauls (21:00 – 03:00) from Trips 2 and 3	0.3176	1573 / 1352	54.85 / 1.36
S_{11}	ω Trip 3 ω hauls less than 3 tonnes total catch weight from Trips 1 and 2	10.92	822 / 681	56.55 / 2.82
S_{12}	ω Trip 3 ω hauls less than 1.5 tonnes total catch weight from Trips 1 and 2	10.8254	1082 / 890	56.73 / 2.47
S_{13}	ω Trip 3 ω night time hauls (21:00 – 03:00) from Trips 1 and 2	0.4819	1390 / 1196	56.93 / 3.16

2. STREAM R-tools (BioSim Tool and SDTool) (Section 1.4.2)







MARE/2020/08

STREAMLINE

FOCUS ON.....

Work Package 2 – Filling information gaps

WP2 Chair I. Bitetto (COISPA); co-Chair S. Kavadas (HCMR)

Partners involved: all partners involved

Core team: P.K. Karachle (HCMR); P. Carpentieri (NISEA); W. Zupa, P. Carbonara, L. Casciaro, M. Donnaloia and M.T. Spedicato (COISPA); V. Raykov, I. Zlateva and M. Yankova (IO-BAS); B. Guijarro, A. Cervantes, J.L. Pérez, E. García and M. Vivas (IEO); I. Thasitis (DFMR); N. Billet (IFREMER); G. Scarcella, F. Grati, F. Masnadi and S. Vitale (CNR); C. Musumeci, A. Massaro, P. Sartor and C. Viva (CIBM); V. Čikeš Keč, B. Zorica, D. Ezgeta Balić, N. Vrgoč and I. Isajlović (IOF)

Based on the existing information and the information gaps identified in WP1, WP2 will generate, in a targeted way, the information and other inputs that are still necessary to design regional work plans for the Mediterranean and Black Sea.
The Tasks of this WP will include the organization of workshops and meetings, as well as case studies to implement or further test the methodologies.

For the success of this work package, MSs will be brought together to collaborate on adc
knowledge and information gaps in a collaborative and coordinated manner.

Project funded by the European Union

MARE/2020/08

STREAMLINE

*Task 2.1 - Sampling design optimization in all the metiers, including SSF, and identification of sampling hierarchy (*I. Bitetto - COISPA)
Partners involved: all partners involved
Core team: C. Musumeci, A. Massaro (CIBM); L. Casciaro and W. Zupa (COISPA); J.L. Pérez, E. García and M. Vivas (IEO); I. Thasitis (DFMR); G. Scarcella and F. Masnadi (CNR); V. Raykov and I. Zlateva (IO-BAS); V. Čikeš Keč and N. Vrgoč (IOF)

The starting point of this task will be the work on the case studies carried out in the STREAM project, involving areas and stocks distributed in the different sub-basins of the Mediterranean and the Black Sea.
The tasks foresees the implementation and testing of additional features in the R tool-box respect to the STREAM project.

The first step of this Task will be thus dedicated to:

a) launch a data call at RCG Med&BS level, similar to the one launched under the STREAM project, to gather updated data in the RCG format to take advantage from a more consolidated dataset (2 more years) to obtain more robust optimization results.

Project funded by the European Union

MARE/2020/08

STREAMLINE

NEW DEVELOPMENTS ON R TOOLS

*Task 2.1 - Sampling design optimization in all the metiers, including SSF, and identification of sampling hierarchy (*I. Bitetto - COISPA)

b) develop additional quality indicators to the ones developed and tested in STREAM taking into account:

- the work carried out in the ICES WKBIOPTIM3
- the Admissible dissimilarity Value (ADV), as a measure of sampling reliability (Wischnewski et al., 2020)

to evaluate the variability of the corresponding relevant estimates (e.g. von Bertalanffy parameters, size at first maturity) and to identify a satisfactory sub-sampling strategy;

Project funded by the European Union



MARE/2020/08

STREAMLINE

NEW DEVELOPMENTS ON R TOOLS

*Task 2.1 - Sampling design optimization in all the metiers, including SSF, and identification of sampling hierarchy (*I. Bitetto - COISPA)

c) develop generalizations of SDTool and BioSim Tool to allow the extraction of samples from the dataset used for the case study, according to combinations of technical, time and spatial characteristics that could be relevant for specific case studies in order to draft a RWP (e.g. Country-GSA).

The sampling design could be set according to a sampling hierarchy (e.g. from individual fish to trip) indicating what sampling levels are included in the multi-stage sampling of the commercial catches and how they are (hierarchically) related to each other, in line with the RDBES concept.

Project funded by the European Union



MARE/2020/08

STREAMLINE

Task 2.1 - Sampling design optimization in all the metiers, including SSF, and identification of sampling hierarchy

First workshop (planned on 29th November- 1st December 2021) for:

- ✓ providing an overview of the methods/approaches adopted in a co-creative process with RCG to fill in the gaps identified in WP1;
- ✓ present the methodology and the new featured R tools to the National experts, allowing them to work on the optimization analysis of biological samplings;
- ✓ sketching out the work and allow the experts to give feedbacks and familiarize with the R tools.

5-6 virtual workshops will be organized with small groups of national experts, for the main Mediterranean geographical basins (western, central, Adriatic, eastern), and the Black Sea:

- ✓ to ease and support the carrying out of the simulations and allow the exploration of the sampling designs considered appropriate and suitable for each specific context;
- ✓ to select, on the basis of the criteria of shared stocks, revise and update, if needed, the case studies presented in the STREAM Final Report;
- ✓ to finalize the relevant STREAM case studies, i.e. the ones useful because dealing with shared stocks, but not included in the STREAM Final Report.

Project funded by the European Union



MARE/2016/22

SDTool OUTCOMES

✓ Perform analyses on sampling optimization in terms of trips

The SD Tool v.2 allows through bootstrap technique to resample the historical data studying the Coefficient of Variation (CV) in association with the number of primary sampling units (e.g. trips) of a given species.

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 term of sampling units) corresponding to that part of the curve is proposed as "optimal" sample size.



MARE/2016/22

SDTool - OUTCOMES

The SD Tool v.2 generalizes the existing SD tool in order to include options allowing a flexible definition of the sampling scheme. The **optimization** can thus be carried out on:

- different technical stratifications introducing options to define the technical strata on the basis of gear (level 4) and/or metier, so grouping strata with similar characteristics;
- different temporal aggregations in order to make flexible the stratification by quarter and/or semester, depending on fisheries and target species specifications;
- data of stocks considered shared among MS in order to get results on the whole area of the stock (not only by GSA)

MARE/2016/22

The SD Tool v.2 generalizes the existing SD tool in order to include options allowing a flexible definition of the sampling scheme. The **optimization** can thus be carried out on:

- different technical stratifications introducing options to define the technical strata on the basis of gear (level 4) and/or metier, so grouping strata with similar characteristics;
- different temporal aggregations in order to make flexible the stratification by quarter and/or semester, depending on fisheries and target species specifications;
- data of stocks considered shared among MS in order to get results on the whole area of the stock (not only by GSA)

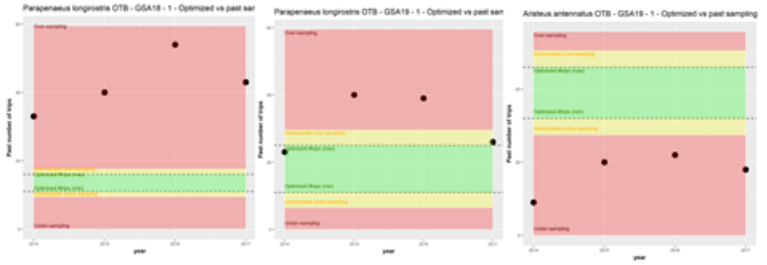
MARE/2016/22

SDTool - outcomes

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

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

	Var1	Var2	Var3	min_trips	max_trips	mean_RR	mean_CV
A. antennatus	GSA18	1	OTB	18	26	0.78	0.14
P. longirostris	GSA18	1	OTB	19	22	0.07	0.22
A. antennatus	GSA19	1	OTB	32	46	0.25	0.13
P. longirostris	GSA19	1	OTB	12	16	0.04	0.23



MARE/2016/22

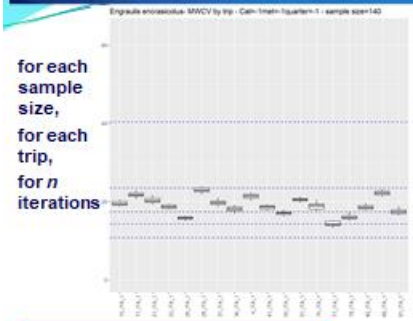
BioSim Tool - OUTCOMES

✓ Perform analyses on sampling optimization in terms of number of individuals to be measured

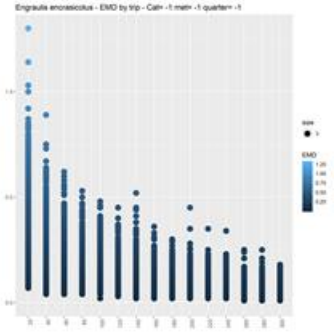
The **BioSim Tool** using bootstrap technique allows to resample the historical data at sample level (i.e. per trip) to identify the minimum no. of fish to be measured in order to avoid oversampling.

This minimum number of individuals to be sampled can be used as a threshold for subsampling.

MARE/2016/22



for each sample size,
for each trip,
for *n* iterations



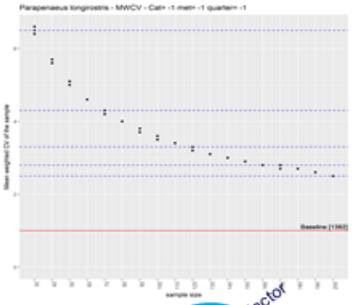
INDICATORS
- mean length,
- **MWCV of lengths**
- median length
- min length
- max length
- number of sampled classes
- number of modes
- **Earth Moving Distance**

MARE/2016/22

BioSim Tool - OUTCOMES

It is possible to assess the impact of changes in the number of length measurements by trip (by commercial category) compared to the baseline (original sample) (variation of each metric compared to the baseline).

samp_size	EMD	MWCV %
30	0.435	6.5
40	0.432	5.6
50	0.435	5
60	0.43	4.6
70	0.436	4.3
80	0.432	4
90	0.433	3.8
100	0.436	3.6
110	0.428	3.4
120	0.438	3.3
130	0.431	3.1
140	0.434	3
150	0.433	2.9
160	0.435	2.8
170	0.431	2.7
180	0.434	2.7
190	0.432	2.6
200	0.438	2.5
Baseline [1382]	-	1



170 Measured individuals → factor

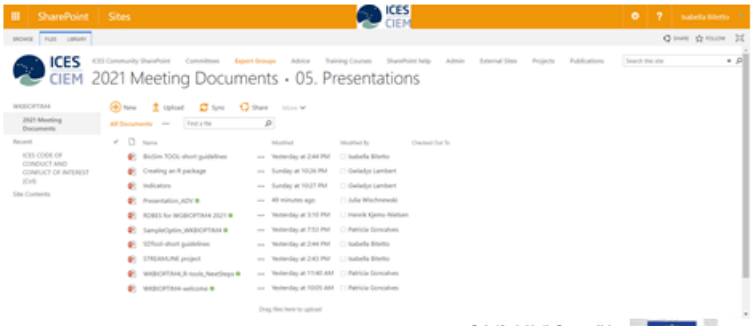
Comm_categ	Sub_sample	Nb_over_Thr	Tot
ALL	1.9	119	170

MARE/2020/08

STREAMLINE

On the sharepoint:

- Short guidelines for SDTool
- Short guidelines for BioSim Tool



Project funded by the European Union

MARE/2020/08

STREAMLINE

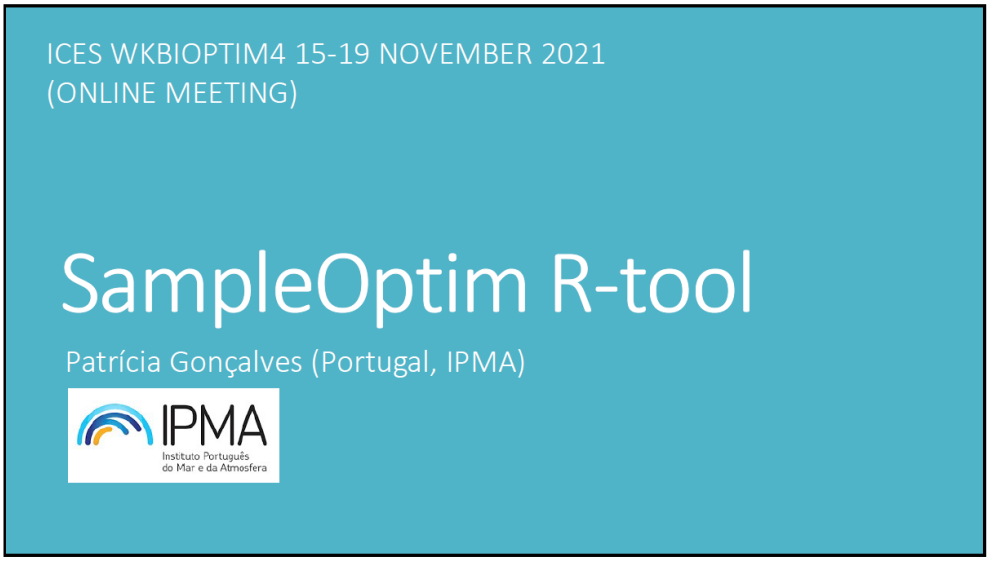
Thanks for your attention!

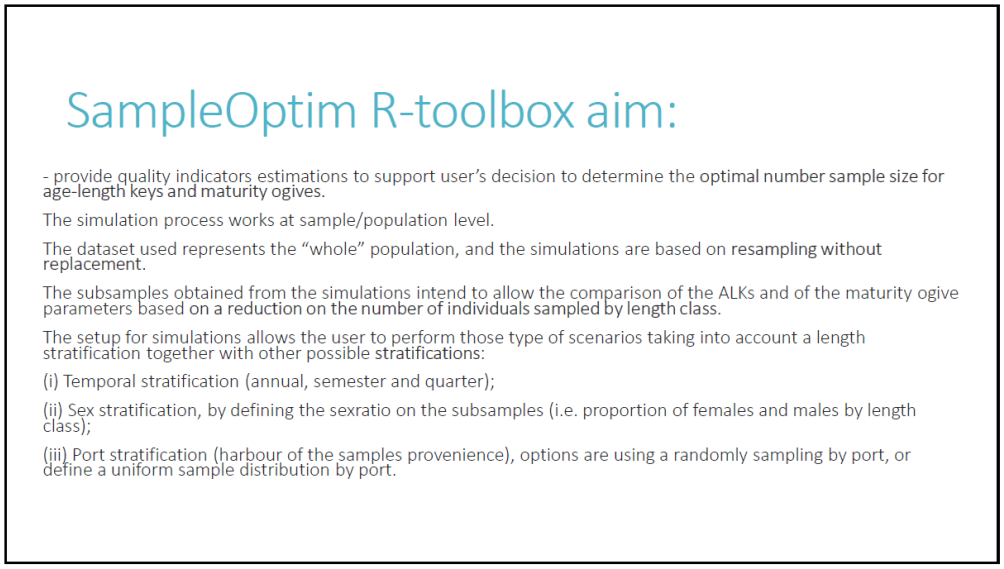
bitetto@coispa.it

COISPA TECNOLOGIA & RICERCA
STAZIONE SPERIMENTALE PER LO STUDIO DELLE RISORSE DEL MARE

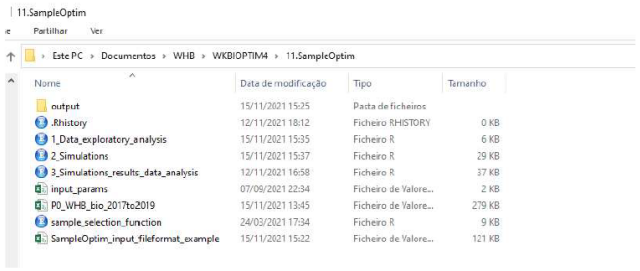
Project funded by the European Union

3. SampleOptim (Section 1.4.3)

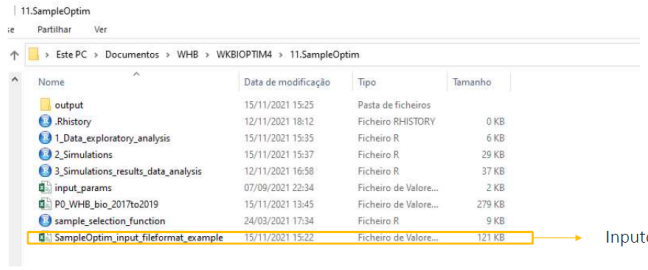




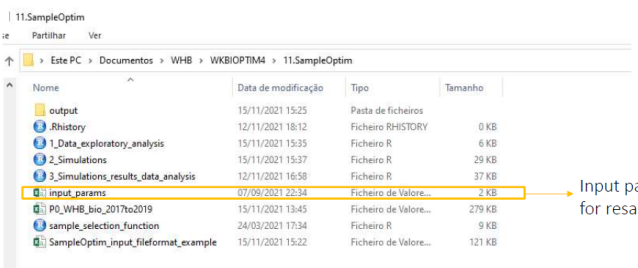
SampleOptim files (R-scripts and .csv)



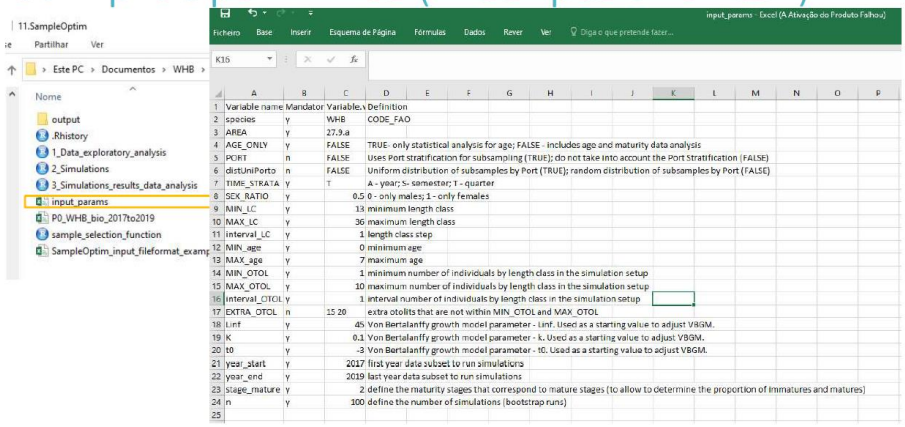
SampleOptim files (R-scripts and .csv)



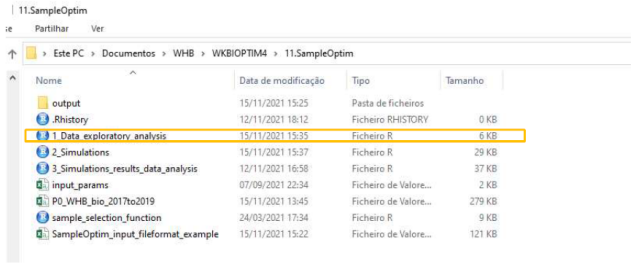
SampleOptim files (R-scripts and .csv)



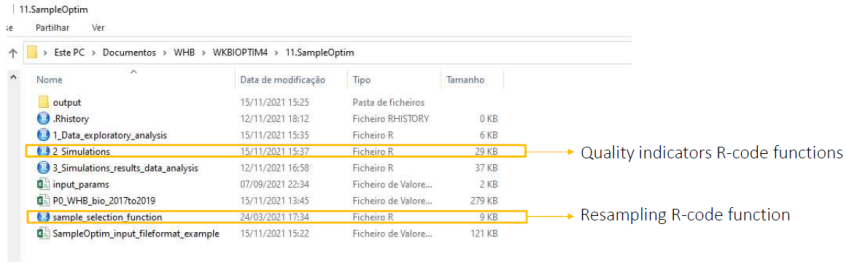
SampleOptim files (R-scripts and .csv)



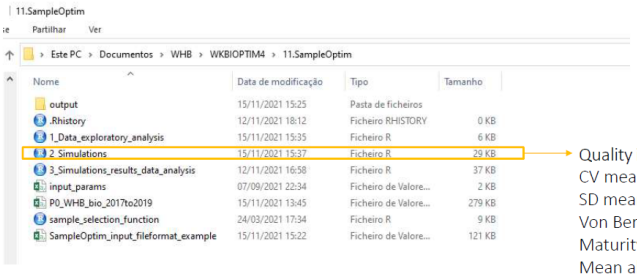
SampleOptim files (R-scripts and .csv)



SampleOptim files (R-scripts and .csv)

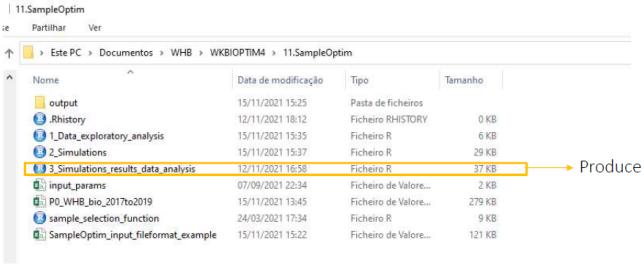


SampleOptim files (R-scripts and .csv)



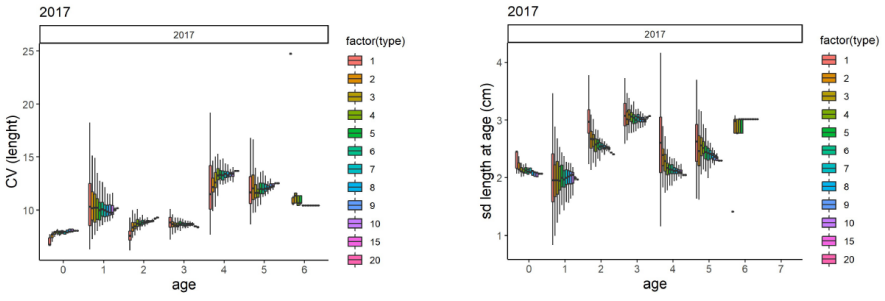
Quality indicators R-code functions:
CV mean length-at-age
SD mean length-at-age
Von Bertalanffy growth parametes (Linf, K, t0)
Maturity parameters (L25, L50, L75)
Mean absolute percentage error (MAPE)
Mean squared percentage error (mspe)
Root mean squared percentage error (RMSPE)

SampleOptim files (R-scripts and .csv)

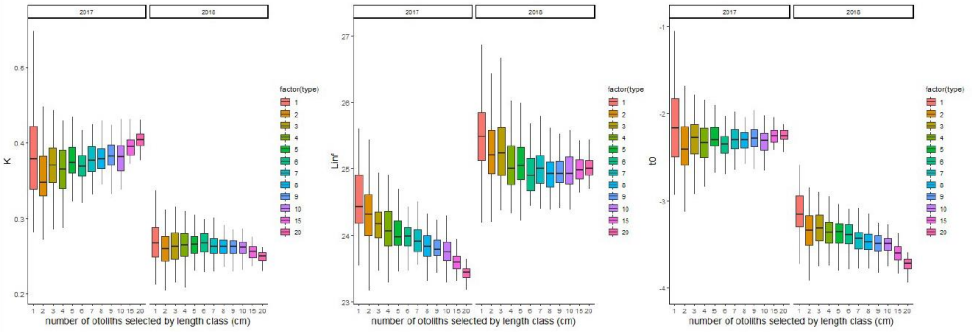


Produce the results

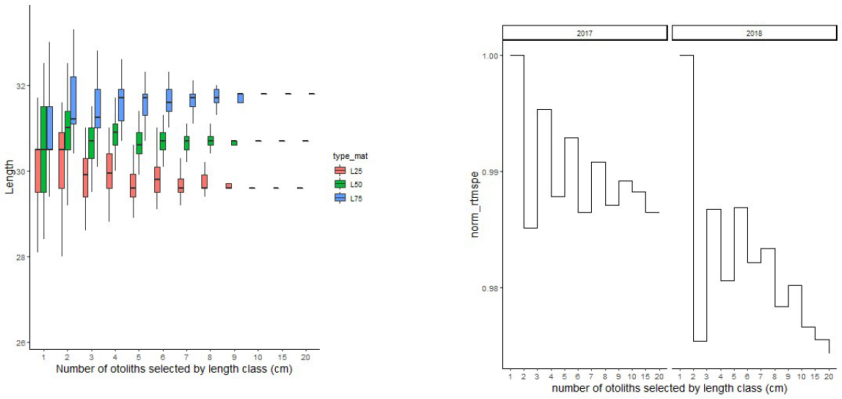
SampleOptim main results

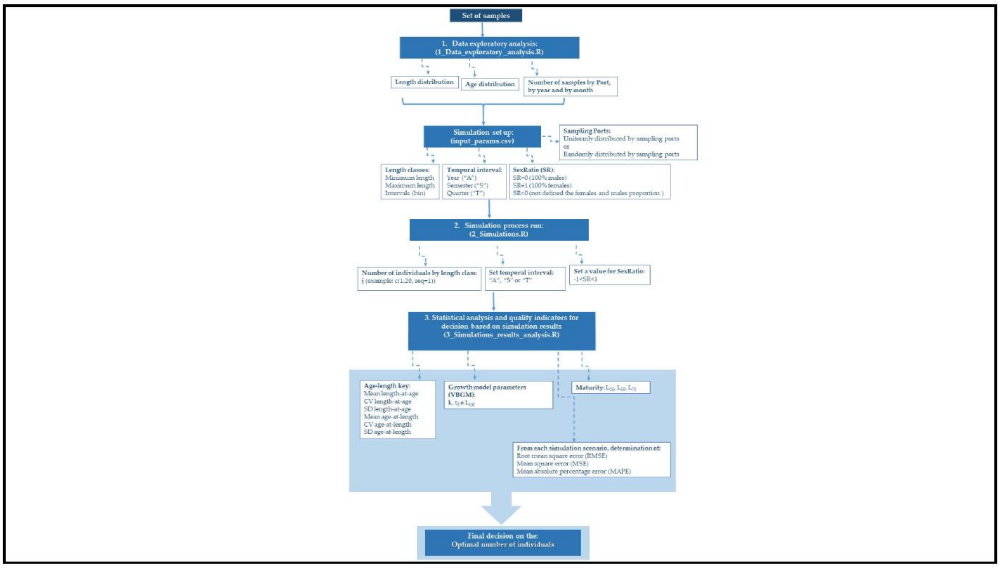


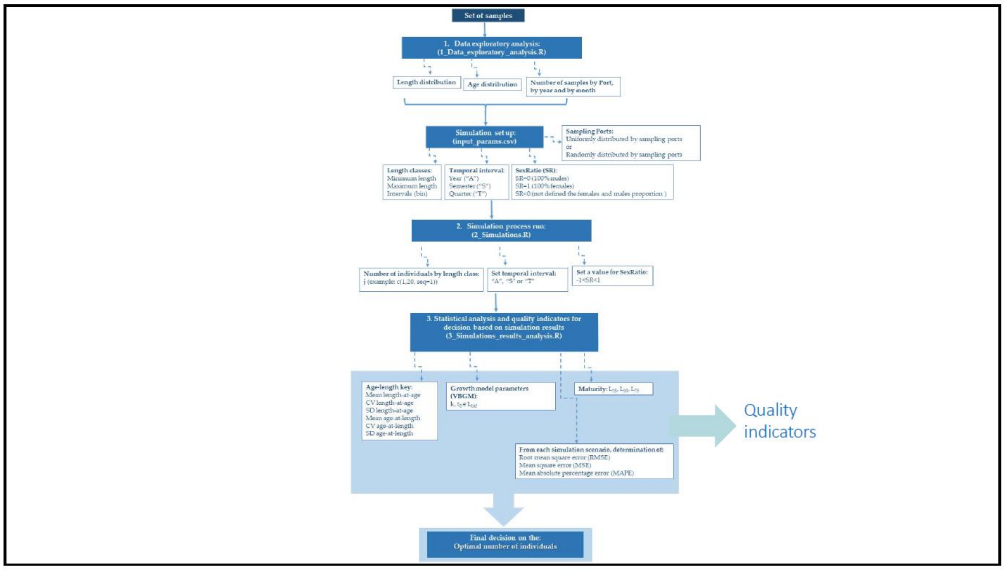
SampleOptim main results



SampleOptim main results






Set of samples
1. Data exploratory analysis:
(1_Data_exploratory_analysis.R)
Length distribution
Age distribution
Number of samples by Port, by year and by month
Simulation set up:
(input_params.csv)
Sampling Ports:
Uniformly distributed by sampling ports
or
Randomly distributed by sampling ports
Length classes:
Maximum length
Maximum length
Intervals (bin)
Temporal interval:
Year ("A")
Semester ("S")
Quarter ("T")
SexRatio (SR):
SR=0 (100% males)
SR=1 (100% females)
SR=0 (input defined the females and males proportion)
2. Simulation process run:
(2_Simulations.R)
Number of individuals by length class:
1 (example: c(1,20, 2eq=2))
Set temporal interval:
"A", "S" or "T"
Set a value for SexRatio:
-1<SR<1
3. Statistical analysis and quality indicators for decision based on simulation results
(3_Simulations_results_analysis.R)
Age-length key:
Mean length-at-age
CV length-at-age
SD length-at-age
Mean age-at-length
CV age-at-length
SD age-at-length
Growth model parameters
(VBGM):
Maturity:
From each simulation scenario, determination of:
Root mean square error (RMSE)
Mean square error (MSE)
Mean absolute percentage error (MAPE)
Quality indicators
Final decision on the:
Optimal number of individuals

Thank you!!!!

4. FishPi4WKBIOPTIM (Section 1.4.4)





Package overview

Simulation of sampling designs

Documentation for package 'FishPi4WKBIOPTIM' version 0.2.0

- DESCRIPTION file

Help Pages

biasPlot	Bias plot
devPlot	Deviance plot
domHist	Frequency distribution of simulated mean lengths
fishPIFormatted_data	Example dataset from FishPi2
formatRDB2FishPi	Format RDB data to run with the FishPi2 simulation package
freqs_plot	Plot simulated length distributions compared to real
freqs_stats	Statistically compare simulated length distributions to real
simul_sampling	Wrapper function to simulate sampling

Wrapper function to simulate sampling

Description

Wrapper function to simulate sampling

Usage

```
simul_sampling(dataset, spp, 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 FormatRDB2FishPi
spp	vector of species names
nsim	number of simulations
psu	options for psu are "fishTripId","vslId","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 "cty_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 (i.e. set to FALSE) - 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.

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



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:

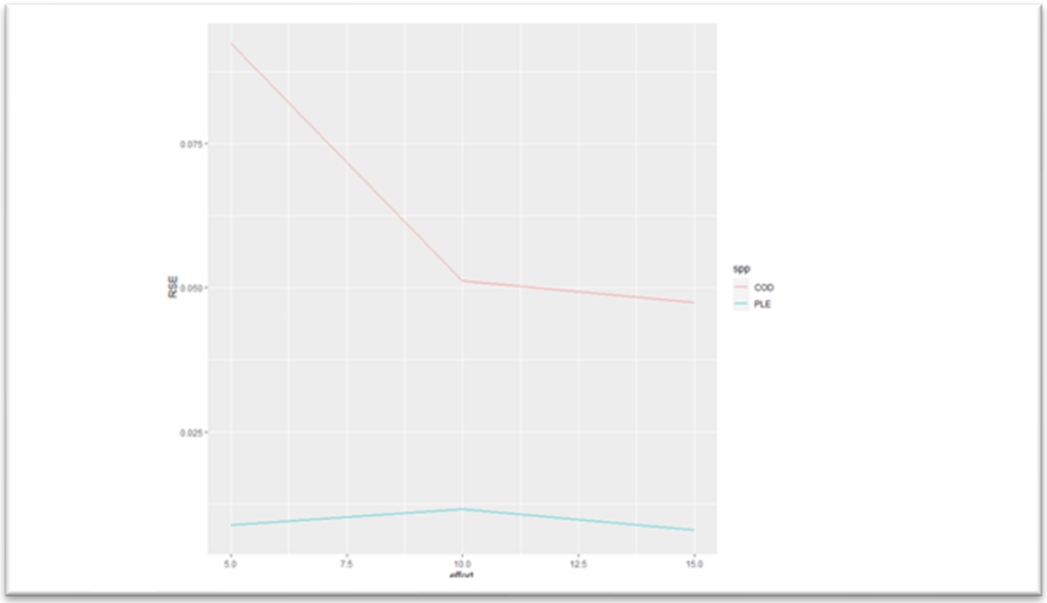
- totBiasEst - NA
- totEst - estimated overall mean length
- sppPop - mean length by species in the population
- RSEest - sd/mean of mean lengths by species (one mean length per simulation)
- sppRSE - mean of RSEest
- meanSppEst - estimated mean length by species
- sppEst - length distributions of each simulation by species
- sppCiLo - upper confidence interval for mean length by species
- sppCiUp - lower confidence interval for mean length by species
- sppSampSize - sample size by species (number of PSUs) - only for "fishTripId" which always either the psu or ssu
- sppBiasEst - bias of the estimate by species, i.e. 100*(meanSppEst/sppPop-1)
- domPop - mean length by domain in the population
- domRSEest - 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)
- domCiUp - uppor confidence interval for mean length by domain
- domCiLo - lower confidence interval for mean length by domain
- domainPop - mean length by domain in the population
- domBiasEst - bias of the estimate by domain, i.e. 100*(meanDomEst/domainPop-1)
- domSampSize - sample size by domain (number of PSUs) - only for "fishTripId" 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

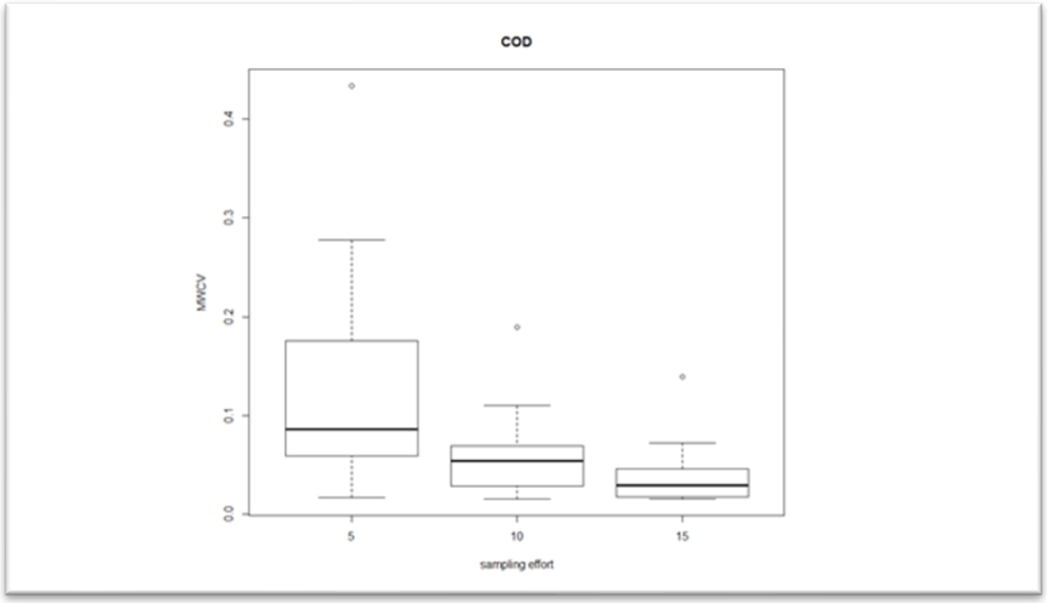
In the second element of the output list, the frequency data at the finest level (i.e. frequencies for a given species at a given sampling effort) contain matrices of simulated data with length classes as rows and simulation number as columns. Each columns sums up to 1 (i.e. fequencies are reported as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

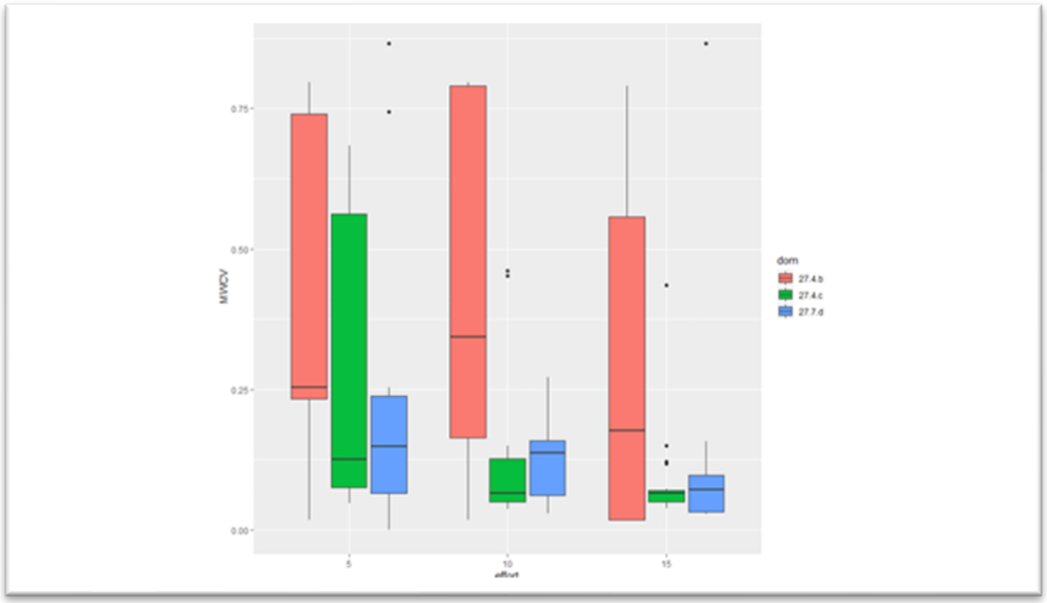
See available example in wrapper function of package

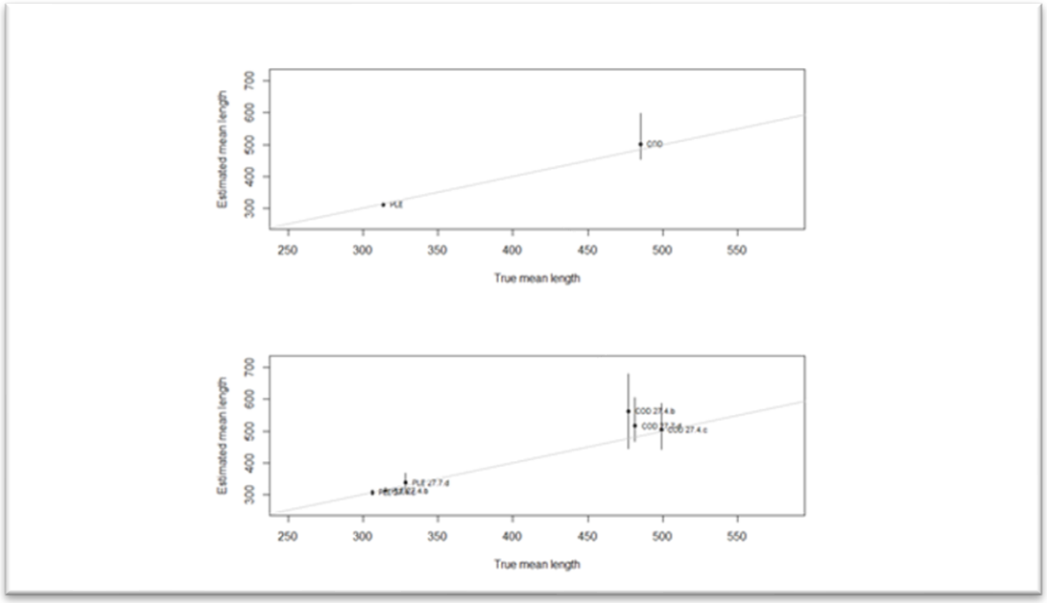
- A number of graphical outputs are available
- Includes MWCV from SampleBioptim and ADV (from 2019 – so not sure if any changes made to method afterwards when publishing paper?)
- All formatted to use package *simul_sampling* outputs that are in convoluted lists formats so not usable functions outside of package

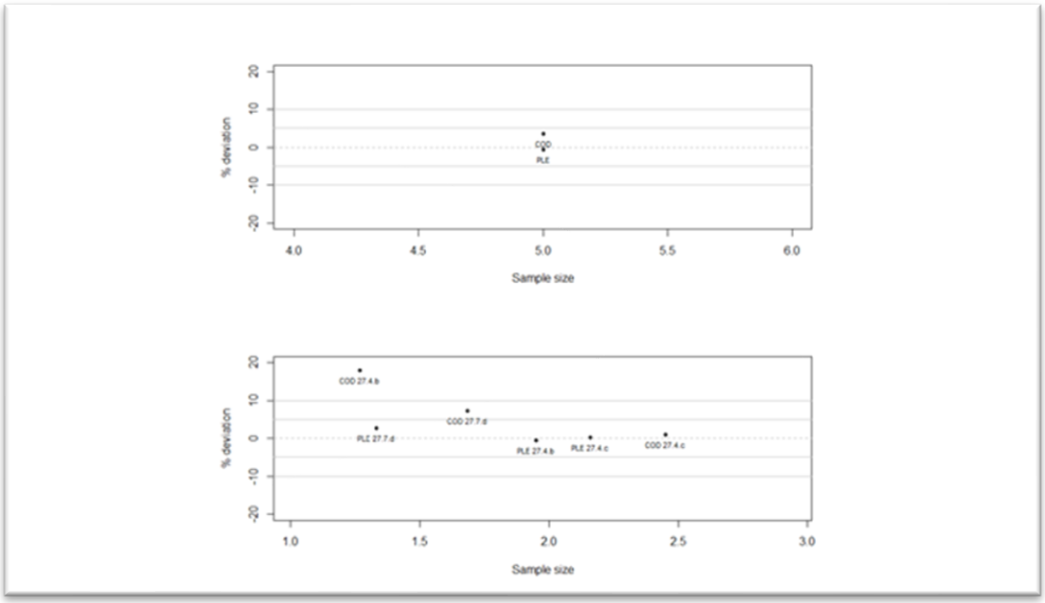
- **Slide below for ideas on list of common indicators and how to present them** (current outputs of FishPi4WKBIOTIM package – these can be retrieved if you run the example in the wrapper function)

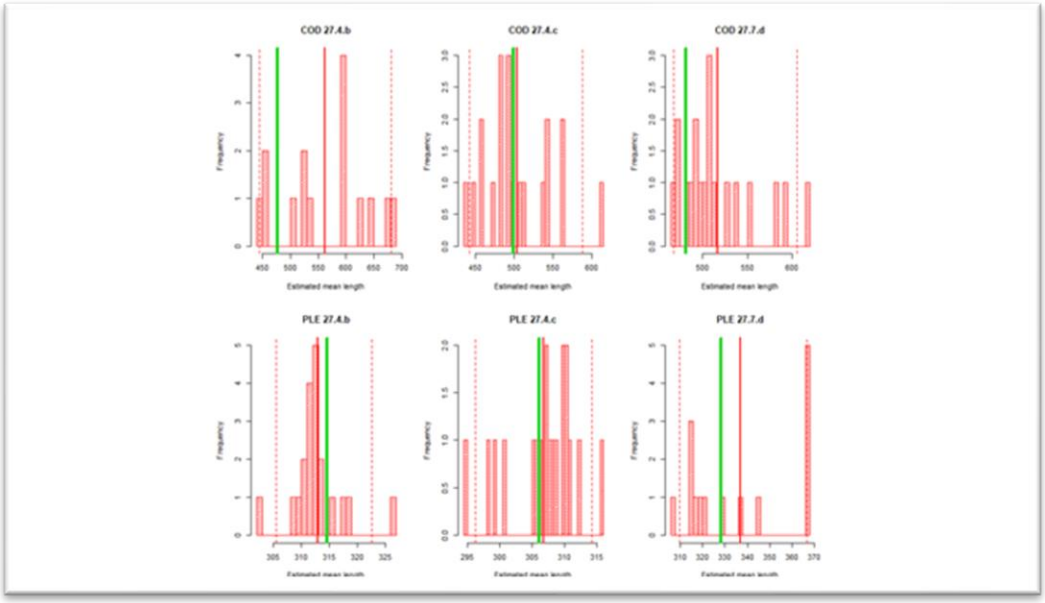

RSE
0.075
0.050
0.025
5.0
7.5
10.0
12.5
15.0
effort
spp
COD
PLE

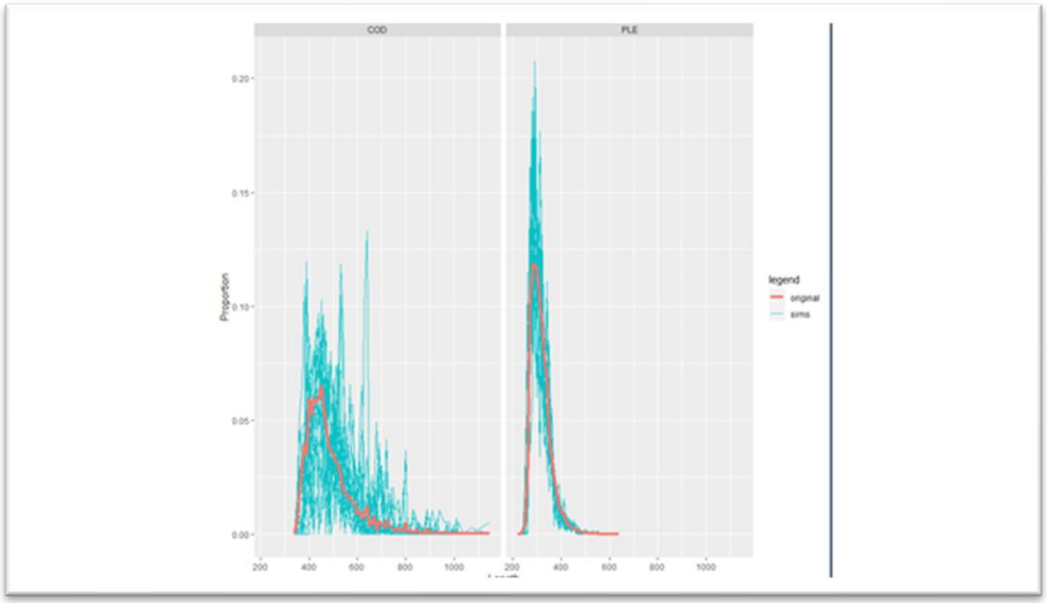

COD
MWCV
0.0
0.1
0.2
0.3
0.4
5
10
15
sampling effort

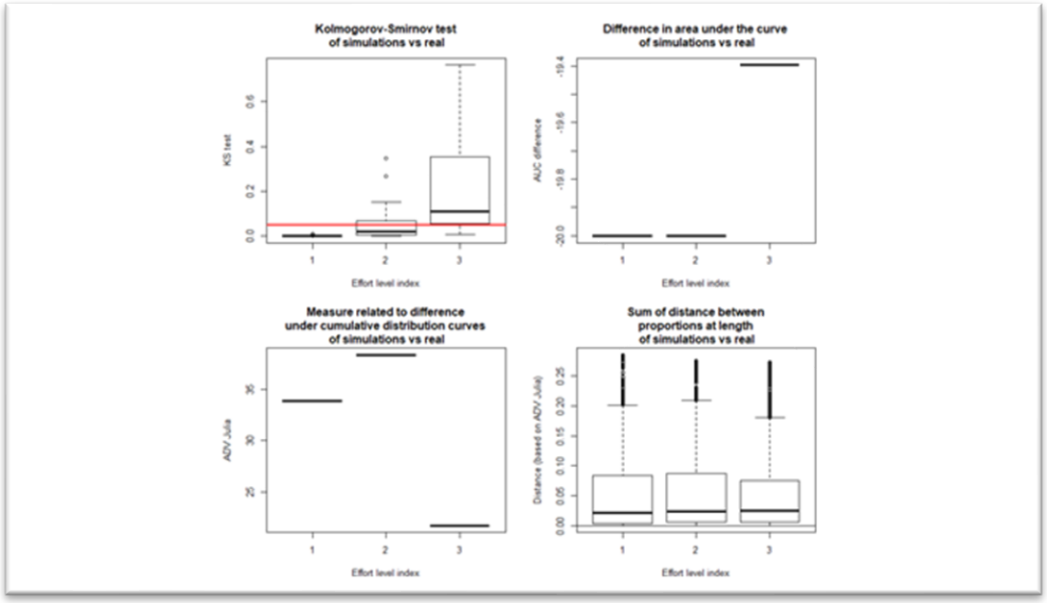

MWCV
0.75
0.50
0.25
0.00
5
10
15
dom
27.4.b
27.4.c
27.7.d


Estimated mean length
300
400
500
600
700
250
300
350
400
450
500
550
True mean length
Estimated mean length
300
400
500
600
700
250
300
350
400
450
500
550
True mean length


% deviation
Sample size
COD
PLE
COD 27.4.b
COD 27.7.d
PLE 27.7.d
PLE 27.4.b
PLE 27.4.c
COD 27.4.c


COD 27.4.b
COD 27.4.c
COD 27.7.d
PLE 27.4.b
PLE 27.4.c
PLE 27.7.d
Frequency
Estimated mean length


COD
PLE
Proportion
legend
original
sims


Kolmogorov-Smirnov test
of simulations vs real
KS test
Effort level index
Difference in area under the curve
of simulations vs real
AUC difference
Effort level index
Measure related to difference
under cumulative distribution curves
of simulations vs real
ADV Julia
Effort level index
Sum of distance between
proportions at length
of simulations vs real
Distance (based on ADV Julia)
Effort level index

Annex 4: R-function to transform RDBES data tables to RDB format

Author: Julia Wischnewski

```
##### Input: RDBES tables and Hierarchy
##### Output: RDB tables - CA, HL, HH and SL
##### At present stage - only for H1 (and H2?), but may be extended
```

```
doRDBEStoRDB <-
function(DEtable=NULL, SDtable=NULL,
VStable=NULL, FTtable=NULL,
FOtable=NULL, SAtable=NULL,
LEtable=NULL, OStable=NULL,
SStable=NULL, TEtable=NULL,
FMtable=NULL, BVtable=NULL,
VDtable=NULL, SLtable=NULL,
selected_upperHierarchy=c(1:16))

{

    SAtable_for_HL <- subset(SAtable, SAlowerHierarchy %in% c("A","B"))
    SAtable_for_CA <- subset(SAtable, SAlowerHierarchy %in% c("A","B","C"))

    if (selected_upperHierarchy %in% c(1))
    {
    FTtable <-  FTtable[,!(colnames(FTtable) %in% c("FOid","OSid","TEid"))]
    SStable <-  SStable[,!(colnames(SStable) %in% c("OSid","TEid","LEid"))]

    DEtable <- subset(DEtable, DEhierarchy==1)


    #### for HL, SL, HH

    FMSAtable <- merge(FMtable, SAtable_for_HL, by=c("SAid"),all.x=TRUE)
    FMSASStable <- merge(FMSAtable, SStable, by=c("SSid"),all.x=TRUE)
    FMSASSFOtable <- merge(FMSASStable, FOtable, by=c("FOid","FTid"),all.x=TRUE)
    FMSASSFOFTtable <- merge(FMSASSFOtable, FTtable, by=c("FTid","SDid"),all.x=TRUE)
    FMSASSFOFTVDtable <- merge(FMSASSFOFTtable, VDtable, by=c("VDid"),all.x=TRUE)
    FMSASSFOFTVDVStable <- merge(FMSASSFOFTVDtable, VStable, by=c("VSid","VDid","SDid"),all.x=TRUE)
    FMSASSFOFTVDVSSDtable <- merge(FMSASSFOFTVDVStable, SDtable, by=c("SDid"),all.x=TRUE)
    FMSASSFOFTVDVSSDDEtable <- merge(FMSASSFOFTVDVSSDtable, DEtable, by=c("DEid"),all.x=TRUE)

    FMSASSFOFTVDVSSDDEtable <- subset(FMSASSFOFTVDVSSDDEtable, DEhierarchy==selected_upperHierarchy)  #### each hierarchy se
quentially?
```

```
    ### determine Landing_country from harbour name - how? ###

    library(devtools)
    #install_github("ices-tools-prod/icesVocab")
    library(icesVocab)

    L_C <- getCodeList("Harbour_LOCODE")

    FMSASSFOFTVDVSSDDE_Location_table <- merge(FMSASSFOFTVDVSSDDEtable,
    L_C[,c("Key","Description","LongDescription")], by.x=c("FTarrivalLocation"),by.y=c("Key"),all.x=TRUE)

    #### for CA

    BVFMSASSFOFTVDVSSDDE_Location_table <- merge(BVtable, FMSASSFOFTVDVSSDDE_Location_table, by=c("SAid","FMid"),all.x=TRUE)

    BVFMSASSFOFTVDVSSDDE_Location_table_age <- subset(BVFMSASSFOFTVDVSSDDE_Location_table, substr(BVtypeMeasured,1,3)=="Ag
e")

    BVFMSASSFOFTVDVSSDDE_Location_table_weight <- subset(BVFMSASSFOFTVDVSSDDE_Location_table, substr(BVtypeMeasured,1,3)=="W
ei")[,c("SAid","FMid",
    "BVnationalUniqueFishId","BVunitName","FMclassMeasured","BVtypeMeasured", "BVvalueMeasured", "BVvalueUnitOrScale", "BVac
curacy","BVmethod")]

    names(BVFMSASSFOFTVDVSSDDE_Location_table_weight)[(ncol(BVFMSASSFOFTVDVSSDDE_Location_table_weight)-4):ncol(BVFMSASSFOFT
VDVSSDDE_Location_table_weight)] <-
    paste0(names(BVFMSASSFOFTVDVSSDDE_Location_table_weight)[(ncol(BVFMSASSFOFTVDVSSDDE_Location_table_weight)-4):ncol(BVFMS
ASSFOFTVDVSSDDE_Location_table_weight)],
    "_weight")

    BVFMSASSFOFTVDVSSDDE_Location_table_maturity <- subset(BVFMSASSFOFTVDVSSDDE_Location_table, substr(BVtypeMeasured,1,3)=
="Mat")[,c("SAid","FMid",
    "BVnationalUniqueFishId","BVunitName","FMclassMeasured","BVtypeMeasured", "BVvalueMeasured", "BVvalueUnitOrScale", "BVac
curacy","BVmethod")]

    names(BVFMSASSFOFTVDVSSDDE_Location_table_maturity)[(ncol(BVFMSASSFOFTVDVSSDDE_Location_table_maturity)-4):ncol(BVFMSASS
FOFTVDVSSDDE_Location_table_maturity)] <-
    paste0(names(BVFMSASSFOFTVDVSSDDE_Location_table_maturity)[(ncol(BVFMSASSFOFTVDVSSDDE_Location_table_maturity)-4):ncol(B
VFMSASSFOFTVDVSSDDE_Location_table_maturity)],
    "_maturity")

    BVFMSASSFOFTVDVSSDDE_Location_table_sex <- subset(BVFMSASSFOFTVDVSSDDE_Location_table, substr(BVtypeMeasured,1,3)=="Se
x")[,c("SAid","FMid",
    "BVnationalUniqueFishId","BVunitName","FMclassMeasured","BVtypeMeasured", "BVvalueMeasured", "BVvalueUnitOrScale", "BVac
curacy","BVmethod")]

    names(BVFMSASSFOFTVDVSSDDE_Location_table_sex)[(ncol(BVFMSASSFOFTVDVSSDDE_Location_table_sex)-4):ncol(BVFMSASSFOFTVDVSSD
DE_Location_table_sex)] <-
    paste0(names(BVFMSASSFOFTVDVSSDDE_Location_table_sex)[(ncol(BVFMSASSFOFTVDVSSDDE_Location_table_sex)-4):ncol(BVFMSASSFOF
TVDVSSDDE_Location_table_sex)],
    "_sex")

        if (nrow(BVFMSASSFOFTVDVSSDDE_Location_table_weight)>0) BVFMSASSFOFTVDVSSDDE_Location_table_age <- merge(BVFMSASSFOF
TVDVSSDDE_Location_table_age,
        BVFMSASSFOFTVDVSSDDE_Location_table_weight, by=c("SAid","FMid",
        "BVnationalUniqueFishId","BVunitName","FMclassMeasured"), all.x=TRUE)

        if (nrow(BVFMSASSFOFTVDVSSDDE_Location_table_maturity)>0) BVFMSASSFOFTVDVSSDDE_Location_table_age <- merge(BVFMSASSF
OFTVDVSSDDE_Location_table_age,
        BVFMSASSFOFTVDVSSDDE_Location_table_maturity, by=c("SAid","FMid",
        "BVnationalUniqueFishId","BVunitName","FMclassMeasured"), all.x=TRUE)

        if (nrow(BVFMSASSFOFTVDVSSDDE_Location_table_sex)>0) BVFMSASSFOFTVDVSSDDE_Location_table_age <- merge(BVFMSASSFOFTVD
VSSDDE_Location_table_age,
        BVFMSASSFOFTVDVSSDDE_Location_table_sex, by=c("SAid","FMid",
        "BVnationalUniqueFishId","BVunitName","FMclassMeasured"), all.x=TRUE)
```

```
    ######################### HL ##################

    HL <- data.frame(matrix(nrow = nrow(FMSASSFOFTVDVSSDDE_Location_table), ncol=length(HL_rdb.names)), stringsAsFactors = F
ALSE)
    colnames(HL) <- HL_rdb.names

    HL$Record_type <- "HL"
    HL$Sampling_type <- FMSASSFOFTVDVSSDDE_Location_table$FTsamplingType
    HL$Landing_country <- substr(FMSASSFOFTVDVSSDDE_Location_table$FTarrivalLocation,1,2)  ## Is it a correct way to extract
a Landing Country?
    HL$Vessel_flag_country <- FMSASSFOFTVDVSSDDE_Location_table$VDflagCountry
    HL$Year <- FMSASSFOFTVDVSSDDE_Location_table$DEyear
    HL$Project <- FMSASSFOFTVDVSSDDE_Location_table$DEsamplingScheme
    HL$Trip_number <- FMSASSFOFTVDVSSDDE_Location_table$FTunitName
    HL$Station_number <- FMSASSFOFTVDVSSDDE_Location_table$FOunitName
    HL$Species <- FMSASSFOFTVDVSSDDE_Location_table$SAspeciesCode
    HL$Catch_category <- FMSASSFOFTVDVSSDDE_Location_table$SAcatchCategory
    HL$Landing_category <- FMSASSFOFTVDVSSDDE_Location_table$SAlandingCategory
    HL$Comm_size_cat_scale <- FMSASSFOFTVDVSSDDE_Location_table$SAcommSizeCatScale
    HL$Comm_size_cat <- FMSASSFOFTVDVSSDDE_Location_table$SAcommSizeCat
    HL$Subsampling_category <- ''   ## I didn't find this field in the RDBES
    HL$Sex <- FMSASSFOFTVDVSSDDE_Location_table$SAsex
    HL$Individual_sex <- ''
    HL$Length_class <- FMSASSFOFTVDVSSDDE_Location_table$FMclassMeasured
    HL$Number_at_length <- FMSASSFOFTVDVSSDDE_Location_table$FMnumberAtUnit
    #HL$Hierarchy <- FMSASSFOFTVDVSSDDE_Location_table$DEhierarchy ## Optional

    HL[is.na(HL)] <- ''
    HL <- aggregate(Number_at_length ~ ., data=HL, function(x) sum(x,na.rm=TRUE))
```

```
    ######################### HH ##################

    HH <- data.frame(matrix(nrow = nrow(FMSASSFOFTVDVSSDDE_Location_table), ncol=length(HH_rdb.names)), stringsAsFactors = F
ALSE)
    colnames(HH) <- HH_rdb.names

    HH$Record_type <- "HH"
    HH$Sampling_type <- FMSASSFOFTVDVSSDDE_Location_table$FTsamplingType
    HH$Landing_country <- substr(FMSASSFOFTVDVSSDDE_Location_table$FTarrivalLocation,1,2)  ## Is it a correct way to extract
a Landing Country?
    HH$Vessel_flag_country <- FMSASSFOFTVDVSSDDE_Location_table$VDflagCountry
    HH$Year <- FMSASSFOFTVDVSSDDE_Location_table$DEyear
    HH$Project <- FMSASSFOFTVDVSSDDE_Location_table$DEsamplingScheme
    HH$Trip_number <- FMSASSFOFTVDVSSDDE_Location_table$FTunitName
    HH$Station_number <- FMSASSFOFTVDVSSDDE_Location_table$FOunitName
    HH$Fishing_validity <- FMSASSFOFTVDVSSDDE_Location_table$FOvalidity
    HH$Aggregation_level <- FMSASSFOFTVDVSSDDE_Location_table$FOaggregationLevel
    HH$Catch_registration <- FMSASSFOFTVDVSSDDE_Location_table$FOcatchReg
    HH$Species_registration <- ''       ## Didn't find what is that in RDBES
    HH$Date <- FMSASSFOFTVDVSSDDE_Location_table$FOstartDate
    HH$Time <- FMSASSFOFTVDVSSDDE_Location_table$FOstartTime
    HH$Fishing_duration <- FMSASSFOFTVDVSSDDE_Location_table$FOduration
    HH$Pos_Start_Lat_dec <- FMSASSFOFTVDVSSDDE_Location_table$FOstartLat
    HH$Pos_Start_Lon_dec <- FMSASSFOFTVDVSSDDE_Location_table$FOstartLon
    HH$Pos_Stop_Lat_dec <- FMSASSFOFTVDVSSDDE_Location_table$FOstopLat
    HH$Pos_Stop_Lon_dec <- FMSASSFOFTVDVSSDDE_Location_table$FOstopLon
    HH$Area <- FMSASSFOFTVDVSSDDE_Location_table$FOarea
    HH$Statistical_rectangle <- FMSASSFOFTVDVSSDDE_Location_table$FOrectangle
    HH$Sub_polygon <- FMSASSFOFTVDVSSDDE_Location_table$FOjurisdictionArea
    HH$Main_fishing_depth <- FMSASSFOFTVDVSSDDE_Location_table$FOfishingDepth
    HH$Main_water_depth <- FMSASSFOFTVDVSSDDE_Location_table$FOwaterDepth
    HH$FAC_National <- FMSASSFOFTVDVSSDDE_Location_table$FOnationalFishingActivity
    HH$FAC_EC_lvl5 <- FMSASSFOFTVDVSSDDE_Location_table$FOmetier5
    HH$FAC_EC_lvl6 <- FMSASSFOFTVDVSSDDE_Location_table$FOmetier6
    HH$Gear_type <- FMSASSFOFTVDVSSDDE_Location_table$FOgear
    HH$Mesh_size <- FMSASSFOFTVDVSSDDE_Location_table$FOmeshSize
    HH$Selection_device <- FMSASSFOFTVDVSSDDE_Location_table$FOselectionDevice
    HH$Mesh_size_selection_device <- FMSASSFOFTVDVSSDDE_Location_table$FOselectionDeviceMeshSize
    #HH$Hierarchy <- FMSASSFOFTVDVSSDDE_Location_table$DEhierarchy      ## Optional

    HH[is.na(HH)] <- ''
    HH <- distinct(HH)
```

```
    ######################### SL ###################

    SL <- data.frame(matrix(nrow = nrow(FMSASSFOFTVDVSSDDE_Location_table), ncol=length(SL_rdb.names)), stringsAsFactors = F
ALSE)
    colnames(SL) <- SL_rdb.names

    SL$Record_type <- "SL"
    SL$Sampling_type <- FMSASSFOFTVDVSSDDE_Location_table$FTsamplingType
    SL$Landing_country <- substr(FMSASSFOFTVDVSSDDE_Location_table$FTarrivalLocation,1,2)  ## Is it a correct way to extract
a Landing Country?
    SL$Vessel_flag_country <- FMSASSFOFTVDVSSDDE_Location_table$VDflagCountry
    SL$Year <- FMSASSFOFTVDVSSDDE_Location_table$DEyear
    SL$Project <- FMSASSFOFTVDVSSDDE_Location_table$DEsamplingScheme
    SL$Trip_number <- FMSASSFOFTVDVSSDDE_Location_table$FTunitName
    SL$Station_number <- FMSASSFOFTVDVSSDDE_Location_table$FOunitName
    SL$Species <- FMSASSFOFTVDVSSDDE_Location_table$SAspeciesCode
    SL$Catch_category <- FMSASSFOFTVDVSSDDE_Location_table$SAcatchCategory
    SL$Landing_category <- FMSASSFOFTVDVSSDDE_Location_table$SAlandingCategory
    SL$Comm_size_cat_scale <- FMSASSFOFTVDVSSDDE_Location_table$SAcommSizeCatScale
    SL$Comm_size_cat <- FMSASSFOFTVDVSSDDE_Location_table$SAcommSizeCat
    SL$Subsampling_category <- ''              ## I didn't find this field in the RDBES
    SL$Sex <- FMSASSFOFTVDVSSDDE_Location_table$SAsex
    SL$Weight <- FMSASSFOFTVDVSSDDE_Location_table$SAtotalWeightLive
    SL$Subsample_weight <- FMSASSFOFTVDVSSDDE_Location_table$SAsampleWeightLive
    SL$Length_code <- FMSASSFOFTVDVSSDDE_Location_table$FMaccuracy
    #SL$Hierarchy <- FMSASSFOFTVDVSSDDE_Location_table$DEhierarchy  ## Optional

    SL[is.na(SL)] <- ''
    SL <- distinct(SL)
```

```
    ######################### CA ###################

    CA <- data.frame(matrix(nrow = nrow(BVFMSASSFOFTVDVSSDDE_Location_table_age), ncol=length(CA_rdb.names)), stringsAsFacto
rs = FALSE)
    colnames(CA) <- CA_rdb.names

    CA$Record_type <- "CA"
    CA$Sampling_type <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FTsamplingType
    CA$Landing_country <- substr(BVFMSASSFOFTVDVSSDDE_Location_table_age$FTarrivalLocation,1,2)  ## Is it a correct way to e
xtract a Landing Country?
    CA$Vessel_flag_country <- BVFMSASSFOFTVDVSSDDE_Location_table_age$VDflagCountry
    CA$Year <- BVFMSASSFOFTVDVSSDDE_Location_table_age$DEyear
    CA$Project <- BVFMSASSFOFTVDVSSDDE_Location_table_age$DEsamplingScheme
    CA$Trip_number <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FTunitName
    CA$Station_number <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FOunitName
    CA$Quarter <- quarter(BVFMSASSFOFTVDVSSDDE_Location_table_age$FOstartDate)
    CA$Month <- month(BVFMSASSFOFTVDVSSDDE_Location_table_age$FOstartDate)
    CA$Species <- BVFMSASSFOFTVDVSSDDE_Location_table_age$SAspeciesCode
    if (!is.null(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_sex))
        CA$Sex <- ifelse(is.na(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_sex),
        BVFMSASSFOFTVDVSSDDE_Location_table_age$SAsex,
        BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_sex) else CA$Sex <- BVFMSASSFOFTVDVSSDDE_Location_table_age
$SAsex
    CA$Catch_category <- BVFMSASSFOFTVDVSSDDE_Location_table_age$SAcatchCategory
    CA$Landing_category <- BVFMSASSFOFTVDVSSDDE_Location_table_age$SAlandingCategory
    CA$Comm_size_cat_scale <- BVFMSASSFOFTVDVSSDDE_Location_table_age$SAcommSizeCatScale
    CA$Comm_size_cat <- BVFMSASSFOFTVDVSSDDE_Location_table_age$SAcommSizeCat
    CA$Stock <- ''                  ## paste0(species_name, area)?
    CA$Area <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FOarea
    CA$Statistical_rectangle <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FOrectangle
    CA$Sub_polygon <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FOjurisdictionArea
    CA$Length_class <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FMclassMeasured
    CA$Age <- BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured
    CA$Single_fish_number <- BVFMSASSFOFTVDVSSDDE_Location_table_age$BVunitName
    CA$Length_code <- BVFMSASSFOFTVDVSSDDE_Location_table_age$FMaccuracy
    CA$Aging_method <- BVFMSASSFOFTVDVSSDDE_Location_table_age$BVmethod
    CA$Age_plus_group <- ''                                         ## ??
    CA$Otolith_weight <- ''                                         ## ??
    CA$Otolith_side <- ''                                           ## ??
    if (!is.null(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_weight))           ## BVvalueMeasured
    CA$Weight <- ifelse(is.na(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_weight),
        NA, BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_weight) else CA$Weight <- NA

    if (!is.null(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVmethod_maturity))        ## BVmethod
        CA$Maturity_staging_method <- ifelse(is.na(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVmethod_maturity),
        NA, BVFMSASSFOFTVDVSSDDE_Location_table_age$BVmethod_maturity) else CA$Maturity_staging_method <- NA

    if (!is.null(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueUnitOrScale_maturity))  ## BVvalueUnitOrScale
        CA$Maturity_scale <- ifelse(is.na(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueUnitOrScale_maturity),
        NA, BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueUnitOrScale_maturity) else CA$Maturity_scale <- NA

    if (!is.null(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_maturity)) ## BVvalueMeasured
        CA$Maturity_stage <- ifelse(is.na(BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_maturity),
        NA, BVFMSASSFOFTVDVSSDDE_Location_table_age$BVvalueMeasured_maturity) else CA$Maturity_stage <- NA

#CA$Hierarchy <- BVFMSASSFOFTVDVSSDDE_Location_table_age$DEhierarchy        ## Optional

    CA[is.na(CA)] <- ''

    } else

{
HL <- NULL
HH <- NULL
SL <- NULL
CA <- NULL
}

return(list(HL = HL, HH = HH, SL = SL, CA = CA))
}
```