SECOND WORKSHOP ON ESTIMATION WITH THE RDBES DATA MODEL (WKRDB-EST2; OUTPUTS FROM 2020 MEETING)

VOLUME 3 | ISSUE 15

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

ICES Scientific Reports

Volume 3 | Issue 15

SECOND WORKSHOP ON ESTIMATION WITH THE RDBES DATA MODEL (WKRDB-EST2; OUTPUTS FROM 2020 MEETING)

Recommended format for purpose of citation:

ICES. 2021. 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

Editors

Kirsten Birch Håkansson • Nuno Prista

Authors

Johnathan Ball • Kirsten Birch Håkansson • Chun Chen • Mary Christman • Liz Clarke • David Currie • Annica de Groote • Jon Elson • Ana Claudia Fernandes • Edvin Fuglebakk • Hans Gerritsen • Josefina Teruel Gómez • Henrik Kjems-Nielsen • Katarzyna Krakówka • Pedro Lino • Richard Meitern • Colin Millar • Karolina Molla Gazi • Duncan Parnell • Nuno Prista • Perttu Rantanen • Petri Sarvamaa • Sven Stoetera • Marta Suska • Julia Wischnewski

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

Contents

i Executive summary

The Second Workshop on Estimation with the RDBES data model (WKRDB-EST2) met online in September 2020 to develop and document R scripts for design-based estimation in the RDBES data model, identify issues in the data model that impact that type of estimation and develop a roadmap for further development of estimation within the RDBES. The main outcomes of WKRDB-EST2 were:

- Discussion and test of a collaborative process, involving all stages of development (from function scripting to package maintenance);
- Collaborative development of data preparation and estimation functions in open-source code (GitHub);
- Carrying out of initial estimation tests that indicated data model suitable for design-based estimation;
- Progress in variance estimation to allow confidence intervals around estimates to be delivered;
- Kick-off of a package "icesRDBES" that will contain the functions, document them, quality check them and make them available to the wider community;
- Discussion of existing data model issues related to estimation and proposal of solutions;

In the final consultation among participants of this WK it was suggested that estimation work in RDBES should be set as a three years' ICES fixed-term working group (WGRDBES-EST). Such WG was considered necessary to secure the intersessional engagement of the participants and a steady and continuous development of all the main estimators relevant for the ICES community including e.g. ratio estimators and procedures for handling of industry refusals. In addition to estimation, the WGRDBES-EST should also reflect on issues such as the long-term maintenance of the code it develops and on a system for peer-review of its work.

ii Expert group information

Expert group name	The Second Workshop on Estimation with the RDBES data model (WKRDB-EST2)
Expert group cycle	Annual
Year cycle started	2020
Reporting year in cycle	1/1
Chairs	Nuno Prista, Sweden
	Kirsten Birch Håkansson, Denmark
Meeting venue and dates	14-18 September 2020, online, 25 participants

1 Introduction

1.1 Overview of RDBES and its development

A short overview of RDBES and its development can be found in the 2019 WKRDB-EST report[1]. Subsequent updates to development strategy, timeline and work-plan are found in the 2020 WGRDBESGOV report (in press).

1.2 Participants and terms of reference for the meeting

The list of participants and the terms of reference of the Second Workshop on Estimation with the RDBES data model (WKRDB-EST2) are presented in Annex 1 and 4, respectively.

Twenty-five participants from 16 institutes and 13 countries attended WKRDB-EST2. Among the institutes present in the WK were institutes from the European Union but also Norway and the UK (ICES countries, outside the EU). An external consultant on statistical estimation of fisheries data was also present in some of the sessions (Mary Christman, USA).

1.3 Agenda and structure of the meeting

The agenda adopted on the first day of WKRDB-EST2 is displayed in Annex 2. To address needs identified during the first WKRDB-EST, a preparatory training course was held that focused on Github, construction of R packages and R-style (08/09; chaired by Colin Millar, David Currie, Kirsten Birch Håkansson and Nuno Prista). Prior to the workshop, a planning meeting was also organized to set the agenda and discuss the work planned for the WK week (10/09; chaired by Kirsten Birch Håkansson and Nuno Prista).

The workshop took place online using MS Teams sessions. In brief, the first plenary of WKRDB-EST2 consisted in an introductory presentation followed by discussion of the ToRs and formation of subgroups. Then participants worked in subgroups for most of the week with plenary sub-group meetings generally taking place each morning. Subgroup chairs met daily at the end of the day to discuss and articulate progress and define the way forward for subgroup work the next day. In the final day, subgroups were asked to reflect on the way forward and on the pros and cons of the online set-up used during the week.

The work was structured in eight subgroups with the following ToRs and subgroup chairs:

SG1 (chaired by Marta Suska and Henrik Kjems-Nielsen):

- Produce functions for data extraction and preparation of datasets
- Adding the probabilities and weights to the datasets

SG2 (chaired by Kirsten Birch-Håkansson):

- Produce function that picks up the prepared tables and creates an estimation object (master table)
- Create functions that run point estimation on estimation object

[1] ICES. 2020. Workshop on Estimation with the RDBES data model (WKRDB-EST; outputs from 2019 meeting). ICES Scientific Reports. 2:5. 106 pp. http://doi.org/10.17895/ices.pub.5956

SG3 (chaired by Liz Clarke):

- Compile and identify variables for variance estimation
- Code a design-based variance estimator for an unbiased estimator of a total (e.g. total catch) under a three-stage sampling design (e.g. vessels, trips, hauls).

SG4 (chaired by Nuno Prista)

- Define structure for Species Selection handling - develop from proof of concept
- Produce function for data preparation of SA table

SG5 (chaired by Edvin Fuglebakk)

- Produce function for univariate or multivariate sample data and estimation
- Trial estimation at sample level

SG6 (chaired by David Currie)

- Set up R-package and think about what it involves (maintenance, test environment, collaboration,…)
- Test inclusion of 1 function in package; incorporate other functions as they get ready

SG7 (no particular chair, met briefly only in last day of workshop)

- Link the subgroups, conduct full estimation

SG 8 (no particular chair, met briefly only in last day of workshop)

- Discuss issues relevant for estimation

Subgroups were autonomous with regards to the organization of their work and reported back to plenary at the end of each day.

Subgroup work was organized in GitHub, https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2. Most of the subgroups used GitHub projects to track progress and issues.

The present report is structured according to the terms of reference of the meeting. First an overview of the RDBES development is given (Section 1); then term of reference a) "Develop and document R scripts for design based estimation for each hierarchy in the RDBES data model" and b) "identify and document any problems with RDBES data model relating to design based estimation" are covered (Section 2 and 3, respectively); finally, the main conclusions of the discussion held on on term of reference c) "Develop a roadmap for future improvements to the estimation procedures within the RDBES" are presented (Section 4).

2 Development and documentation R scripts for design based estimation for each hierarchy in the RDBES data model (ToR a)

2.1 Datasets prepared

Subgroup	Name	Who	Where	Description and comments
subGroup 4	H1_SA_SRSWOR.Rdata	Kasia	SG4/inputs	SA table generated "CreateTestData.R" adapted to sampMathod SRSWOR (no duplicates on species…,)
subGroup 5	input_FMBV_1.rds input_FMBV_Germany_Age.rds	Karolina, Julia	subGroup5/inputs	Modified Lower Hierarchy A (FM, BV) tables from test data - DBErawObj_DK_1966_H1 (Census to SRSWOR, set BVnumTotal to corresponding FMnumAtUnit) and estimated inclusion probability Estimated inclusion probability (Germany's 2018 FM, BV tables/Lower Hierarchy A)
subGroup 1	DK_1965_ESP-AZTI_DCF_Onboard_Sampling_1 DE to BV VD SL	Henrik Kjems-Nielsen	SG1/inputs	testData downloaded from the RDBES (with the IDs included)

2.2 R-code developed

2.2.1 Subgroup 1

Script/function	Who	Language	What it does	Where it can be found	Comments
doDBErawObj.R	**Marta**	**R**	**Reads in all the csv tables (DE.csv, SD.csv,…) from the given folder and if they are consistent with the hierarchy defined in the DE table, produces DBErawObj.rds .**	**https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subGroup1/funs/doDBErawObj.R**	**Works with data model 1.18**
DBEpreparedObj.R	**Johnathan Ball**	**R**	**Reads Rds raw object files and applies probability functiunction, checks input values against calculated values and offers an option to overwrite. DBEraw and DBEprepared objects are retained in the GE, in addition DBEprepared objects**	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subGroup1/funs	**Works with data model 1.18** **Depends on file name following hierarchy structure** **Country_Year_Hierachy.rds**

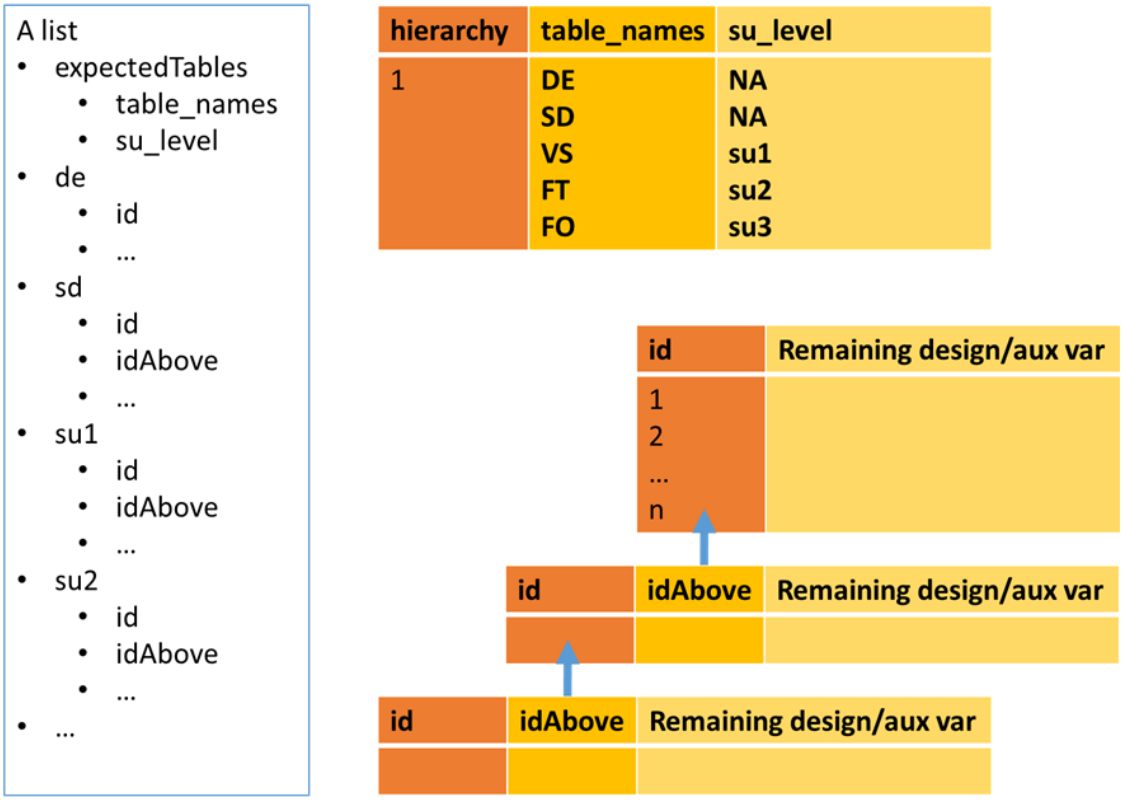
Script/function	Who	Language	What it does	Where it can be found	Comments
			are written out to the output directory **As DBEpreparedObj_Country_Year_Heirachy.rds**		
DBEpreparedObj_2_Clusters.R	**Johnathan Ball**	**R**	**Reads Rds raw object files and applies probability functiunction, checks input values against calculated values and offers an option to overwrite. DBEraw and DBEprepared objects are retained in the GE, in addition DBEprepared objects are written out to the output directory** **As DBEpreparedObj_Country_Year_Heirachy.rds** **Supports Cluster probabilities**	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subGroup1/funs	**Works with data model 1.18** **Depends on file name following hierarchy structure** **Country_Year_Heirachy.rds**
generateProbs_John.r	**Johnathan Ball**	**R**	**Modified version of Nuno's generateProbs function from WKRDB-EST1. Required for DBEpreparedObj.R and DBEpreparedObj_2_Clusters.R** **Possible need to merge with parallel development of original function**	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subGroup1/personal/John	**Works with data model 1.18**
generateClusterProbs_John.r	**Johnathan Ball**	**R**	**Modified version of Nuno's generateProbs function from WKRDB-EST1. Only used on to generate probabilities for Clusters.** **Required for DBEpreparedObj_2_Clusters.R** **Possible need to merge with parallel development of original function**	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subGroup1/personal/John	**Works with data model 1.18** **Needs more testing**

2.2.2 Subgroup 2

Script/function	Who	Language	What it does	Where it can be found	Comments
doDBEestimationObjUpp.R	KBH CC HG	R	Generates the DBE estimation object for the upper hierarchy tables	WK_RDBES/WKRDB-EST2/subGroup2/funs/doDBEestimationObjUpp.R	Works with data model 1.18

This function works on the upper tables of the hierarchy. It takes a list of tables in the prepared data format and identifies the tables that contain the first, second (and so on) sampling units. The output is a list that contains:

- expected_tables: A data frame with information on the hierarchy: which tables correspond to which sampling unit level (e.g, hierarchy 1, su1=VS, su2=FT, su3=FO)
- de: a copy of the design table, this has no information on the sampling units but it does include information on the stratification
- sd: a copy of the sampling details table, necessary for linking to the hierarchy tables.
- su1: a dataframe with a selection of columns from the table that contains the primary sampling unit, so for hierarchy 1, this would be the VS table, for hierarchy 2 the FT table etc.
- su2: the secondary sampling unit.
- su… until the n'th sampling unit.



Summary of discussions and decisions:

- Hierarchy does not need to be an argument to the function, it can be inferred from the DE table.
- Only the columns relevant for estimation are returned in the su objects. (Information on stratification, clustering, probabilities as well as Id columns necessary for linking to other tables.
- Variable names in the su tables are stripped of the first 2 characters, so VSincProb becomes incProb. This should make it easier for generic estimation scripts to work.
- Id columns that are foreign keys are not stripped of the first 2 characters in order to maintain the link to auxiliary tables.
- The id column that is the primary key will be stripped of its first 2 characers (it will now be simply: "id"). An additional column will be added: idAbove (*maybe rename to parentId)*. This should make it easy to link the tables in the correct hierarchy.
- The function was tested for all hierarchies. It failed for a number of hierarchies and it needs to be checked if this is an input data problem or a bigger issue.
- There are a number of outstanding issues, which are reported on in Section 2.3.2.

2.2.3 Subgroup 3

Script/function	Who	Language	What it does	Where it can be found	Comments
estimateHT-multistag-eSRSWOR	**Liz Clarke, Sven Stötera, Annica de Groote**	**R**	**Uses a Horvitz-Thompson estimator to calculate point and variance estimates for a univariate population total in a multi-stage sampling design with SRSWOR (or census) in each stage.** **This will also work for a single stage design.**	**Soon to be found in:** **https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/sub-Group3/funs**	**This currently uses Kirsten's test sample data format, after some edits.** **The calculations have not yet been checked. This will be done over the next week.**

Summary of discussions and decisions:

- We started with the problem of estimating the variance of a Horvitz-Thompson estimator of a total under a multi-stage sampling design.
- We documented the estimation formulae partly for a general design with sampling without-replacement in each stage, partly for the specific case of simple random sampling without replacement (srswor) in each stage (Annex 5). The function is restricted to the specific case of srswor.
- We focused on estimation of a population total since this is a parameter that can be of real interest (for instance, total weight of discards) whilst the variance is easier to estimate than for instance the variance of an estimator of a ratio.
- We calculated the variance of the overall total in one step instead of by table (stage) in order to make sure that all contributions to variance were considered.
- We noted that joint inclusion probabilities are not, at this point, included in the database. In general, those joint probabilities are needed for variance estimation if a without-replacement design is used. For the special case of srswor this was however not a problem, everything we needed for estimation was there.

- The function we developed is specific for a three-stage design with srswor in each stage and includes both point and variance estimation.
- We considered trying bootstrap for variance estimation as well but since this would be more complicated and time-consuming we left that for another time.
- We documented the estimation formulae for a multi-stage design using with-replacement sampling in the first stage (Annex 5). This approach substantially simplifies the variance estimation. The estimation formulae rest however on the assumption that if a primary stage unit is selected more than once, it is independently subsampled as many times as it is drawn. In practice, if the primary sampling units are vessels, they may not be subsampled again if they are selected more than once.
- We did not code the with-replacement design since we ran out of time.

2.2.4 Subgroup 4

Script/function	Who	Language	What it does	Where it can be found	Comments
CreateCommSpeciesTable.R	Kasia Krakówka	R	Generates zero (new SA for commercial species) and new table with sample values of commercial species. The examples of Lophiidae.	https://github.com/ices-eg/WK_RDBES/blob/master/WKRDB-EST2/subgroup4/personal/Kasia/CreateCommSpeciesTable.R	For this function data has been prepared manually (SA, SL)
generateZerosInSA.R	Nuno Prista et al.	R	Generates zeros for census of species in SL	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subgroup4/funs	
newSAquery	Nuno Prista et al.	R	Generate zeros for a query of a spp	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subgroup4/funs	
simFreqOccfromSLsampling,R	Nuno Prista	R	Simulates sampling from a species list	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subgroup4/funs	
mainScript.R	Nuno Prista et al.	R	demonstrates use of generateZerosInSA in adding zeros and missing values to SA table based on SL information & demonstrates a query of a species to the SA table (comparing it with what was reported in SL)	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2	
findDiffObsTyp.R	Ana Fernandes	R	Checks for NA in 'observationType' and if different 'observationType' are present in the same FO	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subgroup4/funs	

Script/function	Who	Language	What it does	Where it can be found	Comments
check-SampMeth.R	**Ana Fernandes**	**R**	**Checks if the sampling method in the SS table is different from CENSUS. Condition for creating the zeros**	**https://github.com/ices-eg/WK_RDBES/tree/master/WKRDBEST2/subgroup4/funs**	
checkCluster.R	**Ana Fernandes**	**R**	**Checks if there is clustering in SS table**	**https://github.com/ices-eg/WK_RDBES/tree/master/WKRDBEST2/subgroup4/funs**	**This function can be applied for checks in other tables**
checkStratif.R	**Ana Fernandes**	**R**	**Checks if there is stratification in SS table**	**https://github.com/ices-eg/WK_RDBES/tree/master/WKRDBEST2/subgroup4/funs**	**This function can be applied for checks in other tables**

2.2.5 Subgroup 5

Script/function	Who	Language	What it does	Where it can be found	Comments
doDBEestimationObjLow_List.R	**Karolina**	**R**	**Creates estimation object for multiple samples**	SubGroup 5 folder	**This script is an adaptation to multiple samples of Edvin's doDBEestimationobjLow.R for a single sample. Merge two scripts?**
computeDBEresultsTotalPointLow.R	**Edvin**	**R**	**Estimates totals for a single SA**	SubGroup 5 folder	**Works on a preliminary definition of the estimation object. Need to be adapted to the structure used by doDBEestimationObjLow_List.R**

2.2.6 Subgroup 6

Script/function	Who	Language	What it does	Where it can be found	Comments
icesRDBES	David Currie, Richard Meitern, and Petri Sarvamaa	R	Draft package	https://github.com/ices-eg/WK_RDBES/tree/SG6/WKRDB-EST2/subGroup6/icesRDBES	The draft package incorporates a few draft functions from the other groups, but not all of them. Some tests for these functions have been defined.
.travis.yml	Richard Meitern	YAML	At each commit to the branch where the file exists checks if the package code is OK for submitting to CRAN	https://github.com/ices-eg/WK_RDBES/blob/SG6/.travis.yml	If the file is moved to package specific repro the line 4 before_install: … should be removed

2.2.7 Subgroup 7

Script/function	Who	Language	What it does	Where it can be found	Comments
doDBEestimationObjUppMid.R	Kirsten	R	Generates the DBE estimation object for the upper and middle hierarchy tables	https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subGroup7/funs	
estimationSuggestions.R	Edvin	R	Examples of estimation for H1 totals using external packages like the 'survey' package, input is from a preliminary version of the estimation function at the SA level.	https://github.com/ices-eg/WK_RDBES/tree/e9c26fbdd1e6b6167b53fc09e3290192adae7352	

2.3 R-code to be done and ideas left to be developed

A lot of the subgroups used GitHub to track ideas, leftovers and issues, https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2

2.3.1 Subgroup 1

Script/function	Who	What is missing, what would be nice to develop
doDBErawObj.R	Marta Suska	Quality checks (to be discussed, depends on what will be checked in the RDBES) Now the function can load the files from one Sampling Design at once. Could be developed, so that it works for more complicated designs (multiple combinations of Country, Year, Hierarchy, SamplingDesign inside one DE) Check how it works for the hierarchies with 'mandatory tables not in the hierarchy' or optional tables
DBEpreparedObj.R & DBEpreparedObj_2_Clusters.R	Johnathan Ball	Consider altering dependance on filename to values in table (general rewrite)
generateProbs_John.r	Johnathan Ball	Merge with other versions developed by other subgroups Add Roxygen documentation Add support for probunits Add support for SWSWOR Rename once finalised
generateClusterProbs_John.r	Johnathan Ball	Add Roxygen documentation Add support for probunits Add support for SWSWOR Rename once finalised

Some of these functions are dependent on Davis Curries work with the RDBES format, https://raw.githubusercontent.com/davidcurrie2001/MI_RDBES_ExchangeFiles/master/RDBES_Functions.R. To avoid this dependency, then it would be beneficial to add the functions to icesRDBES.

2.3.2 Subgroup 2

Script/function	Who	What is missing, what would be nice to develop
doDBEestimationObjUpp.R		**Handling of tables that are not a part of the sampling hierarchy** The function does not deal with mandatory or optional tables that are not a part of the sampling hierarchy, so presently a 'lazy' version that needs a foreign key. Information from these tables may be needed for estimation (particularly domain estimation). The IDs are all in the output object (as long as they are in the input object) so it may not be necessary to include them. The function needs to check if the optional tables are present or not, and handle the linking accordingly. The function needs to be tested with data downloaded from the RDBES, so all the needed id's are there. Presently the functions have been tested with data, which haven't been downloaded from the RDBES and therefor some of the id's may be missing. **Auxiliary data** How and when to include relevant auxially data should be considered

Ideas about the estimation process - The estimation process could be as follows

1. Prepare data, subset relevant data, eg. remove invalid hauls, remove catch components that are not of interest etc.
2. doDBEestimantionObjUpp
3. ComputeDBEresultsTotalPointUpp - (not developed yet). Arguments:
    a) parameter. What do you want to estimate? Pick a relevant column name, for the upper hierarchy, this could be "FOdur" if you want fishing hours, for lower hierarchy it could be "FMnumberAtUnit" etc. This conveniently identifies the table as well as the parameter
    b) level. At which level do you want to estimate your parameter? E.g. by trip, vessel, stratum etc. This can be specified by naming the relevant table, e.g. 'FT' for trip, 'TE' for quarter, month etc.
    c) estimator. "HT" or "HH"
    d) estimationObj, the estimation object (from DBEestimantionObjUpp)

Summary of discussions and decisions:

- The group considered having an argument to specify what to estimate that would simply be a string like "Fishing duration per trip". A reference table could then be used to link the substring "Fishing duration" to FOdur in the FO table and "per trip" to the "FT" table, but it was considered more robust to ask the user to specify the parameter and level directly (as above, e.g. "FOdur" and "FT")
- It may be necessary to subset the input data, e.g. only discards unless an additional argument is included to specify the catch fraction. Are there other issues like that?

Unresolved questions

- Should the estimation be done in a stepwise way (separately for each level in the hierarchy) or all in one go.
- Should the function allow different estimators for different levels in the hierarchy (e.g. HT for su1 and HH for su2?

2.3.3 Subgroup 3

Other functions that would be nice to develop

- Functions for estimating variance are missing for many (most) possible sampling designs. We suggest moving forward case by case; that is, developing variance estimation functions for the most common sampling designs first and not try to deal with everything at once (that can easily become an overwhelming task). What designs are most common need to be agreed upon.

2.3.4 Subgroup 4

Script/function	Who	What is missing, what would be nice to develop
CreateCommSpeciesTable.R	Kasia Krakówka	It would be nice to develop example with sprat and herring

Other functions that would be nice to develop

- The handling of situations where commercial species are sampled that involve mixtures of biological species needs review.

2.3.5 Subgroup 6

Script/function	Who	What is missing, what would be nice to develop
icesRDBES		All functions from the other sub-groups will need incorporating, and tests written for them.

2.3.6 Subgroup 7

Script/function	Who	What is missing, what would be nice to develop
doDBEestimantionObjUppMid.R		Proper integration with the work done in subgroup 4 Proper testing Proper integration with the work done in subgroup 5

Where we are

At this early stage of development is has been beneficial to develop bits of pieces of the big puzzle in smaller subgroups. This setup allowed the handling of certain issues that relating to specific tables in the hierarchies e.g. infinite levels of subsampling in the sample table.

The idea with subgroup 7 was to link work done in the other subgroups and develop a preliminary process for estimation. Some of the developed functions need to be merged before a full estimation object (upper, middle and lower) can be made; other functions are more standalone and just need to find a good place in the overall process; lastly, some of the functions are more alike test of concept and may therefore not need incorporation.

The group started to combine functions from the upper and middle hierarchy. The estimation object for these parts started to take shape, but much more work is need (see table above). Further, the subgroup started to map the developed functions within the overall process, from RDBES output to final estimates, see Figure 2.3.6-1. Progress on the latter mapping was a bit hindered by time available and the fact that a lot of the functions developed in the subgroups still are at an early-stage of development and need further work and integration at the subgroup level before being incorporated in the map.

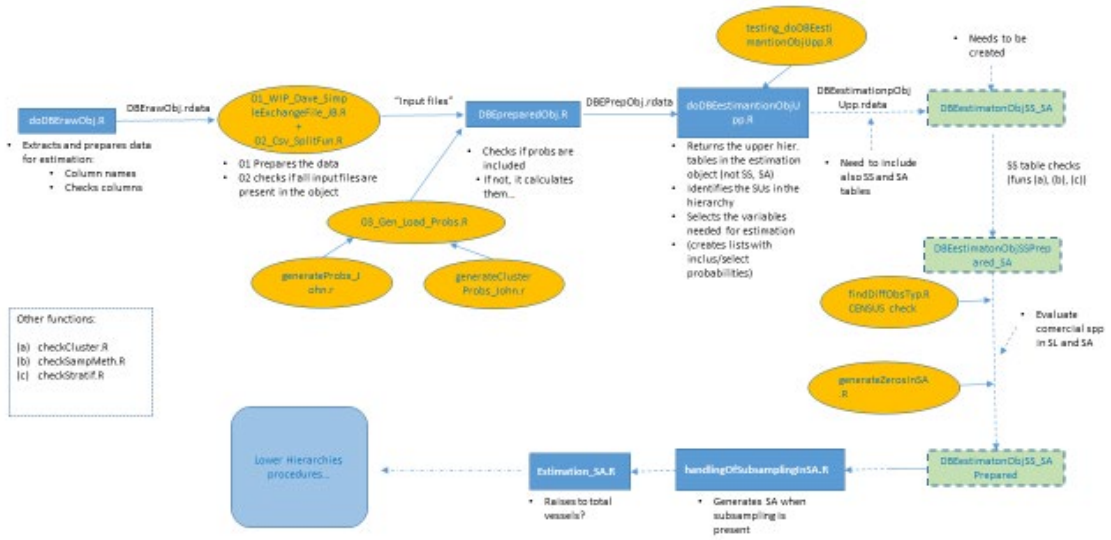


Figure 2.3.6-1. 1st attempt to map the functions developed at WKRDB-EST and WKRDB-EST2. The map should be considered very preliminary.

Next steps

- Agree on the setup: is the better option still a generic estimation object as suggested at WKRDB-EST? if so, what will be its format and how we will adapt all relevant functions to it? if not, what alternatives can be suggested?
- Continue the mapping of all the functions. It will be beneficial to do this at an early stage, so code being developed by different individuals does not drift too much apart.
- Continue documenting the overall work flow from RDBES outputs to final estimates.

2.4 Other developments

2.4.1 Subgroup 3

We documented the formulas needed to make design-based point and variance estimation for two cases of three stage sampling: sampling without replacement in all stages, and sampling with replacement in stage one. See Annex 5.

2.4.2 Subgroup 4

Code was developed that simulates the estimation of frequency of occurrence for programmes that involve the sampling of random species from a frame of positive landings (= landings containing the species)[2].

[2] see simFreqOccfromSLsampling.R in https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST2/subgroup4

2.4.3 Subgroup 6

Proposal for community development of RDBES package ("icesRDBES")

The aim of these guidelines has been not to make package development too onerous for the wider ICES community whilst also not putting too large a workload on the package maintainers – we also still need to ensure a minimum standard is met (e.g. the code is valid and in a consistent style, verifiably does what it's supposed to, and can be submitted to CRAN).

Once the guidelines are agreed we should create a simple guideline document that can be given to potential contributors describing what they need to do. These contribution guidelines can be included in the repo e.g. https://github.com/tidyverse/dplyr/blob/master/.github/CONTRIBUTING.md

- The package source code is hosted within its own repo by ICES e.g. in ices-tools-dev / ices-tools-prod
- A small group of maintainers will need to be volunteered, including someone from the Secretariat
- There are two branches within the repo: master and dev
    - The master branch is protected so that only the maintainers can commit to it
    - The dev branch is used for all development work – contributors can commit directly to it
    - A pull request needs to be created when we want to merge the development branch into the master branch – the maintainers will need to approve the pull request
    - Release labels can be applied to the master branch to keep track of releases
    - A "lint" tool is configured that will compare committed code to a defined style and warn if there any problems – contributors should endeavor to resolve any issues flagged
    - Contributors need to be given commit access to the repo – ICES have a work-flow for this.
    - Contributors should pull the dev branch, make and commit changes on their local machine, and only push changes back to GitHub once their work is consistent (e.g. a new function is created and documented)
    - It is preferable for contributors to only use base R but the following packages (and their dependencies) are also allowed: data.table, and dplyr (not the whole tidyverse). If contributors wish/need to use other packages this must be discussed beforehand.
- Before any development is started an "issue" should be raised on the GitHub repo – this can be to point out a bug in existing code, improve existing functions, or describe new functions that are required.
    - The proposed changes can then be discussed and agreed with the maintainers and other relevant people – this should also act as peer-review system for the statistical content of the proposed development
    - All contributed developments should be linked to an issue – they will not be included in a future pull request if they are not.
    - If a bug has been identified, then working code to demonstrate that bug should be provided in the issue – this can then be converted into a test within the package. The package code should then develop to a point where the test can be passed.
- The fastest way for contributors to get their code included in the package is to provide code that fully meets our package standards. These are:

  - The contributor is using the latest version of R, roxygen and any dependencies
  - For each function an R file exists in the R directory
  - Full roxygen2 documentation has been generated for that function and any data included. A good example of comprehensive documentation can be provided by the "gam" function (https://www.rdocumentation.org/packages/mgcv/versions/1.8-33/topics/gam). This level of documentation might not be appropriate or feasible in all cases but does show some important features to bear in mind.
  - The contributor has defined tests for all new functionality
  - Devtools::check has been run successfully on their local machine
  - The code passes the automatic lint checks
- We recognize that not every contributor will be able to meet this standard so have also defined a minimum standard – only supplying code at this level will result in it taking longer to be included in the package. The minimum standard of contribution is:
  - The contributor is using the latest version of R, roxygen and any dependencies
  - For each function an R file exists in the R directory
  - The Roxygen2 documentation comments have been generated for that function and any data – the descriptions of functions and parameters should be written. The Roxygen2 function should have a short description – see the documentation for "exampleFunction" for an example.
  - Some simple examples of using the functions and its expected outputs are supplied
  - The code doesn't have any major issues raised by the lint checks
- Periodically the maintainers will update the package in CRAN. If people want/need the latest version of the package it can always be installed directly from GitHub.

Automation of testing and style checking on GitHub using Travis CI

(Richard)

The github repro of the package should be configured so that after any commit to the master branch the package is checked if it meets CRAN submission requirements. An example of a ".travis.yml" file for performing such checks for the current and next R release contains the following:

```
language: r
before_install: cd  WKRDB-EST2/subGroup6/icesRDBES
sudo: false
cache: packages
warnings_are_errors: true
matrix:
 include:
 - r: release
 - r: devel
   latex: false


# repository
repos:
 CRAN: https://cloud.r-project.org


r_packages:
 - rmarkdown
 - covr
```

```
 - testthat

r_github_packages:
 - jimhester/lintr
 - ices-tools-prod/icesVocab

after_success:
 - Rscript -e 'covr::codecov()'
 - Rscript -e 'lintr::lint_package(linters=lintr::with_defaults(object_name_linter=lintr::object_name_linter(styles = "camelCase")))'
```

For detailed ".travis.yml" documentation see https://docs.travis-ci.com/user/languages/r

This ".travis.yml" file needs to be present only in the master branch and is not needed for dev or other branches. This will ensure that the checks are run only after commit to the master and will therefore avoid unnecessary checks and emailing. If the package is moved to the main directory of the repro the line pointing to the package location (i.e before_install: cd WKRDB-EST2/subGroup6/icesRDBES) should be removed.

The package also checks for code style using the lintr package
(i.e. - Rscript -e 'lintr::lint_package(linters = lintr::with_defaults(object_name_linter=lintr::object_name_linter(styles = "camelCase")))')

and test coverage (i.e. - Rscript -e 'covr::codecov()').

Any linter errors will be added as commit comments for each commit in the github repro. Also these errors will be sent to the committer email and to all others watching the github repository.

To enable Travis Continuous Integration (CI) automated checks this functionality must be enabled from the github marketplace by the owner of the github repository. For the https://github.com/ices-eg/WK_RDBES repository it has been enabled and the status of Travis CI checks for each commit can be accessed at travis-ci.org/github/ices-eg/WK_RDBES

Automated running of lint and style checks using pre-commit hooks

It is possible to automatically run lintr and styler checks on your local machine before committing code. This isn't a requirement for people to contribute code but some contributors will find it useful. The full instructions are found at the following address: https://github.com/lorenzwalthert/precommit#installation

A summary of the steps are:

1. Check that python3 is installed and install if necessary
2. Run pip install
   `pip3 install pre-commit --user`
3. Install R precommit package
   `install.packages("precommit")`
4. Run at the root of the git repository (change cwd in R)
   `library(precommit)`

5. precommit::use_precommit()
6. Running git commit should run the various checks automatically. See the config file https://github.com/ices-eg/WK_RDBES/blob/SG6/.pre-commit-config.yaml for all the checks.

How to easily style your R code and identify problems

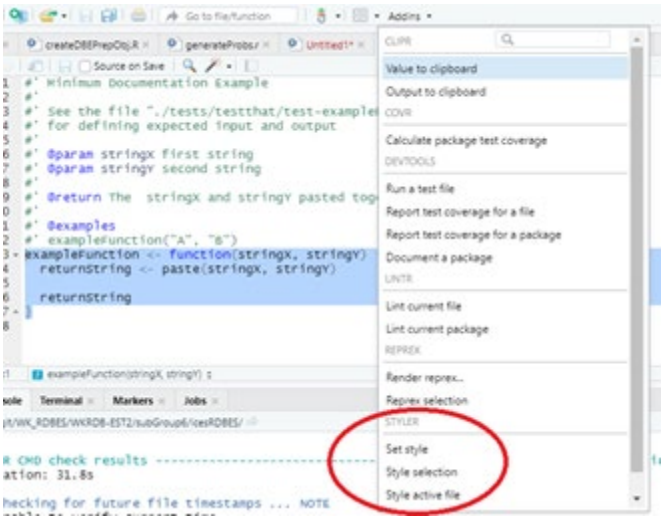
Following the style guidance can be burdensome so here are some instructions to make your life easier using two R packages: **styler** and **lintr**

The first step is to install the R packages styler and lintr (they are both available on CRAN)

Styler

First we'll use the styler package to fix some of the simple style problems in our code. The default style for styler is the tidyverse style – this will be fine for us. We can run styler either on a selection of code or a file, as you prefer.

If you are using RStudio then when you install styler an RStudio add-in is created – this allows you to easily style sections of text, or the active file. Just select the text, and then choose "Style selection" from Addins:



You can also just run styler from the console e.g. run the following to style a file called myFunction.R:

styler::style_file('myFunction.R")

Note that the file you run styler on will be changed without being backed-up

Lintr

This is a tool to help identify and remove problems in your code – in particular we are using it to identify any remaining style problems.

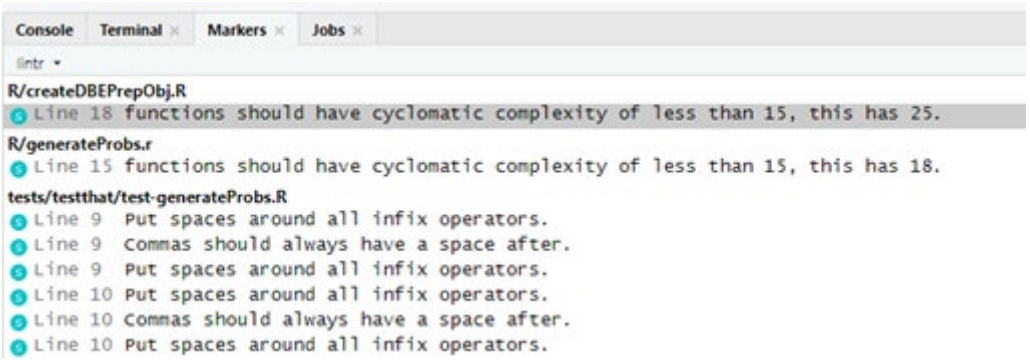
Run the following command to identify problems in a file called myFunction.R:

lintr::lint("myFunction.R",linters=lintr::with_defaults(object_name_linter=lintr::object_name_linter(styles = "camelCase")))

You can also run lintr on a whole package:

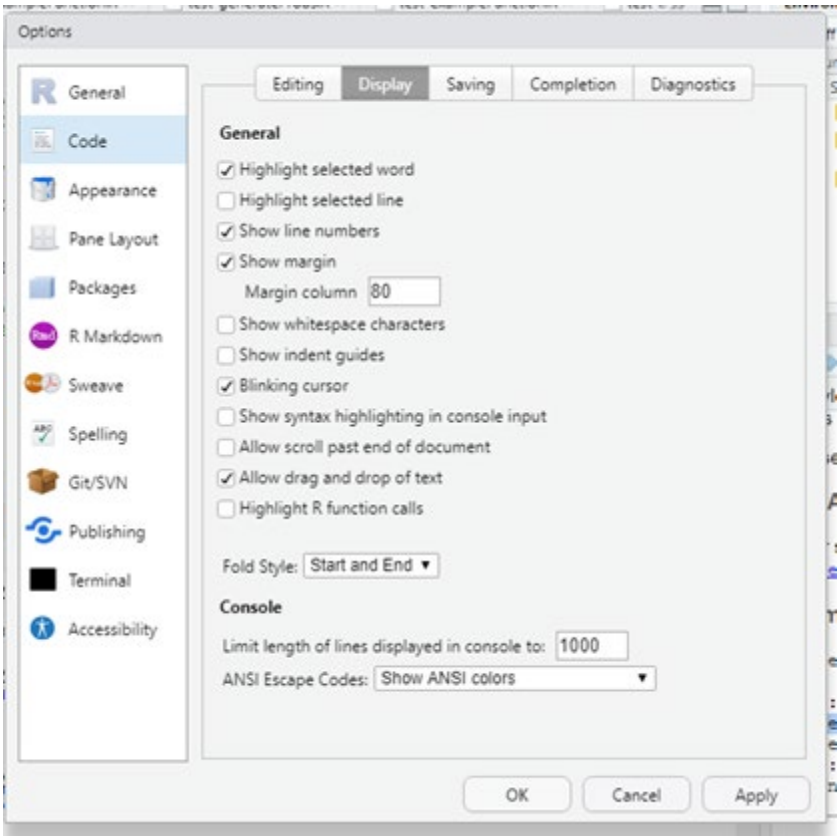
lintr::lint_package(linters=lintr::with_defaults(object_name_linter=lintr::object_name_linter(styles = "camelCase")))

If you run the lintr command in RStudio you will see a "Markers" pane open to show any problems:



If you have already run styler on your code then most of the easy problems should have already been resolved – you will probably just be left with the problems that you need to manually fix. Click on the entry in the Marker window and you will be taken to the line that is causing a problem.

In RStudio to help with fixing lines that are too long you can choose to show a margin at 80 characters:



2.4.4 Subgroup 7

See section 2.3.6

3 Identify and document issues problems with RDBES data model relating to design based estimation (ToR b)

During the workshop, an online document was made available to subgroups working under ToR a) where they documented difficulties experienced in the use of the RDBES data model for design-based estimation. Using this document as a basis, a final issue list to be considered under ToR b) was produced (Table 3-1). The table contains both the issues detected in the subgroups and pending issues related to design-based estimation in the RDBES GitHub. It is worth mentioning that the list of issues analysed in WKRDB-EST2 is comparatively small in relation to similar issue lists reported in WKRDB-EST and WKRDB-POP and WKRDB-POP2. The reduction in the number of issues over time is a reflection of the maturation of the RDBES data model with regards to the upload issues thus far report, but it is important to mention that thus far only a very limited number of practical estimation trials have been done using the RDBES data. It is expected that the issue list will once again grow in size as this last component is addressed.

Table 3-1 displays the details of issues discussed in WKRDB-EST2. Time constraints, the online nature of the workshop and a decision taken to focus participant time in the wrapping up of developments achieved under ToR a) made it impossible to discuss all issues. Detailed examples and analysis of the addressed issues are presented in Annex 6. Non-addressed issues are only described in Table 3-1**. It must be noted that Annex 6 does not represent definitive conclusions of WKRDB-EST2 on any of the issues, but rather reflections and suggestions of the EG that should be further considered by the core-group of development of RDBES and, if necessary, by a future edition of the present WK.**

Table 3-1 List of issues related to design-based estimation discussed in WKRDB-EST2. Issues 1-6 were discussed during the WK (see suggestions of improvements needed in Annex 6). Perceived importance for design-based estimation (DBE) is also reported.

#	Origin	Brief Description	Perceived Importance for DBE [1 = Low; 2 = Medium; 3 = High]	Status
1	GitHub Issue lists	In the current data structure it is not possible to identify what part of the population is not covered by sampling (a.k.a. "out-of-frame"). Lack of coverage can take place at any level of the hierarchy and when not documented may be the cause of biases in estimation, e.g., 1. a country does not include smaller vessels in its sampling frame for the VS table; 2. a country has a sampling scheme targeting 2 specific size categories and only these 2 size categories are declared in its SA table; 4. a country only takes samples from one area out of several areas fished in a given trip; 3. a country randomly samples 48 out of 52 weeks in the year Main issue **https://github.com/ices-taf/RDBES_Core_Group/issues/9** Related issues: **https://github.com/ices-tools-dev/RDBES/issues/52** **https://github.com/ices-tools-dev/RDBES/issues/15**	3	Discussed in subgroup, presented in plenary. Basis for further discussion in the core-group of development of RDBES in Annex 6.
2	GitHub Issue lists and sub-Group 1 of WKRDB-EST	Now it's possible to submit the data with several different selection methods within each stratum. Is it statistically ok? What if there is e.g. SRSWR and NPAH method in one stratum. Should numberTotal and numberSampled include only the units sampled probabilistically or all the units that were sampled in this stratum (as is said in the description of the column in the data model). How are we going to carry out all the estimation in this case? Related to **https://github.com/ices-tools-dev/RDBES/issues/71**	3	Discussed in subgroup, presented in plenary. Basis for further discussion in the core-group of development of RDBES in Annex 6.
3	Subgroup 3 of WKRDB-EST2	Joint inclusion probabilities are required for estimation of variance for unequal probability designs. (In simpler cases e.g. SRSWOR, SRSWR, these can be calculated from sample size and population size.) These are not currently incorporated into the RDBES format and would require either repetition of rows or matrices of joint inclusion probabilities for units within a sample. We propose that these are not incorporated into the RDBES but that institutes requiring these more complicated analyses import them into R for the estimation in a separate format, or use other imported information to calculate them, as required. Look into: **https://github.com/ices-tools-dev/RDBES/issues/76**	1	Discussed in meeting of subgroup chairs and plenary. Core-group recommended to insert guidelines in documentation. Initial thoughts on such guidelines in Annex 6.
4	Subgroup 4 of WKRDB-EST2	Missing instruction in documentation: Under concurrent sampling, SSuseCalcZero should be used like a quality indicator - when data submitter reports "Yes", concurrent sampling can be considered finished and zeros calculated. When the data submitter reports	1	Discussed in subgroup, presented in plenary. Core-group recommended to insert guidelines in

#	Origin	Brief Description	Perceived Importance for DBE [1 = Low; 2 = Medium; 3 = High]	Status
		"No", concurrent sampling was not finished (note: SSnumberTotal should be reported NA) and zeros should not be calculated for any of the missing species.		documentation. Initial thoughts on such guidelines in Annex 6.
5	Issue lists	Missing instruction on how to declare species*size combinations in the SA table. There are several alternatives possible – see details in https://github.com/ices-tools-dev/RDBES/issues/70	2	Discussed in subgroup, presented in plenary. Initial thoughts on pros and cons of different alternatives present in Annex 6. Basis for further discussion in the core-group of development of RDBES in Annex 6.
6	Issues List	Minutes as sampling units **https://github.com/ices-tools-dev/RDBES/issues/74#issuecomment-693637148**	2	No conclusion during WKRDB-EST2, see Annex 6.
7	Issues List	Representative fish **https://github.com/ices-tools-dev/RDBES/issues/7** Related to: **https://github.com/ices-tools-dev/RDBES/issues/10** **https://github.com/ices-tools-dev/RDBES/issues/21** Other notes: see discussions in skype core-group 2020-06-10	1	Not discussed during WKRDB-EST2.
8	Issues List	Quota sampling and non-probabilistic sampling	3	Not discussed during WKRDB-EST2.
9	Issues List	Is the SA table working with regards to BMS in terms of generating the correct 0s and NAs	3	Not discussed during WKRDB-EST2.

4 Develop a roadmap for future improvements to the estimation procedures within the RDBES (ToR c)

In the final day of the workshop, participants were split into their subgroups and asked to address the following questions:

- Road-map forward
- What are the next step in your subgroup?
- What can be achieved by means of Intersessional Development?
- Do we need a Workshop Next year?

In addition, considering the special circumstances of this workshop (held online due to the covid pandemic) the following questions were also asked:

- What worked for you and what didn't this week?
- Did the subgroup size work? Better larger? Better…individual?!
- Did the combination plenaries/subgroups, communication via subgroup chairs work for you?
- Did the distance set-up work for you? Not the same but…better? Worse?

Answers given by the different subgroups are displayed in Table 4-1 and 4-2.

In what concerns the way forward in terms of development of estimation within the RDBES, participants signalled their wish to finalize their work by means of some kind of intersessional work or dedicated time to estimation within the activities of the core group of development of the RDBES. They also expressed their wish to continue developing code that facilitates ratio estimation within the RDBES. Workload at national institutes is however significant: most participants estimating they would be able to allocate a maximum of 1 afternoon per month to intersessional development. There was support for the continuation of the work in a new meeting the following year and a general agreement that, under the present time limitations, the work to be done will necessarily be longer-term and better fit to the 3-year work-plan of an ICES WG than to one-off WKs like WKRDB-EST or WKRDB-EST2. Participation in a WG would also facilitate availability of time for intersessional work at national institutes and its coordination by WG chairs.

In what concerns workshop format, participants highlighted some of the pros and cons of the present online format. Overall, the type of work carried out under the WK, which essentially involves the coordination of code-development in small groups of participants, was found suitable for the purpose at hand and online participation. Significant challenges were however identified in the online format, including higher level of disturbances from other work, technology failures, lack of overview, and the absence of human interactions during problem solving.

It is the chair's opinion that the move from a WK to a WG is the most adequate way to sustain the longer term development needs of ICES in terms of estimation within the RDBES. The online format used this year was productive and suitable to the work being done. It also seems to have allowed the participation of a wider number of participants than a physical meeting. Online participation is not however a silver-bullet for development needs of the RDBES - it is sometime difficult to secure the participants and engaged and the longer term planning perspective, engagement and mutual support tends to fall short when the meeting is held online.

Table 4-1 Results of the questionnaire on way forward in terms improvements to the estimation procedures within the RDBES

	What is the next step?	What can be achieved by means of Intersessional Development?	Do we need a Workshop Next year?
SG 1	Compile work done this week, continue discussing issues related to estimation	1 afternoon per month - it is ok, but there is preparation time so it may be difficult to always be prepared	It is important to continue work, a lot still to be done and discussed
SG 2	If estimation could continue as ISSG similar to the core group, that would be the best way. It may be too much work for some. Possibility might be to have the core group monthly and in weeks between have the ISSG working. So, less specification of data model and more focus on estimation	Quite a bit. 1 afternoon per month is too little	Yes, mainly because gets people closer for a continuous period, and there is a report on status
SG 3	Quality checks should be possible to look up manually at certain stages and sections even with all auto quality control put in	---	Mixed response - maybe ok, maybe strapped with time
SG 4	Functions work but not finished. Want to finish and test functions with real data	1 afternoon per month? Achievable amount of time but not during data calls, so maybe better a day every two months or a larger meeting every 6 months.	Yes
SG 5	Putting SG work together, making sure the 1st 2 steps (database extract, prep of estim object) Addressing the issue list, including issues that were discussed this workshop and are still to have a final decision	CL and CE Part are still to be considered. Could be developed intersession. Refine the code and not let the code die because of lack of "maintenance". 1 afternoon per month	Yes, that hopefully focus on the most complex cases. All the loose ends from this year. Would be nice to have solved the simple cases ahead of it.

Table 4-2 Results of the questionnaire on workshop format

	What worked for you and what didn't? (online nature)	Did the subgroup size work? Better larger? Better…individual?!	Did the combination plenaries/subgroups, communication via subgroup chairs work for you?	Did the distance set-up work for you? Not the same but…better? Worse?
SG 1	Worked well overall but it easy to be requested by other work and family	It is ok if 2-3 people are fully dedicated otherwise if someone misses you work alone and discussions are not so productive	It was fine	Yes, when not disturbed by other things. But is nice to be face to face.
SG 2	Ok size. Some people leave the SG because it is online. Communication worked very well also because we were few. It could have been more productive if the connections between SG were clearer from the start	---	---	Physical is preferred. Avoids work and family distractions during WK time
SG 3	WK gets more done, less probably is we split it	---	---	---

	What worked for you and what didn't? (online nature)	Did the subgroup size work? Better larger? Better…individual?!	Did the combination plenaries/subgroups, communication via subgroup chairs work for you?	Did the distance set-up work for you? Not the same but…better? Worse?
SG 4	---	Nice size (n=3) but hard from distance when people cannot be there Not so much time to meet other people face to face Worked well in terms of writing code, much easier at home than with in a busy room with others Enjoyed discussing code with other group members	---	Technology does not always work. Not so much time to meet other people face to face. Worked well in terms of writing code, much easier at home than with in a busy room with others - emails always come in anyway
SG 5	Actual coding, we have results, unlike in EST1 where we mostly discussed formats, more basic survey stat issues, etc. Connection problems, waiting in the lobby, etc	Nice in small groups - easier to cooperate and task sharing Prep meeting was ok to set up the SG ahead of meeting 3-4 was a good size	Yes	Easy to work individual and concentrate than in a room Negative: Connection problems, waiting in the lobby, distraction from other tasks and co-workers, etc.

Annex 1: List of participants

Name	Institute	Country (of institute)	Email
Ana Claudia Fernandes	IPMA	Portugal	acfernandes@ipma.pt
Annica de Groote	Swedish University Agricultural Sciences	Sweden	annica.isaksson.de.groote@slu.se
Chun Chen	Wageningen University & Research	The Netherlands	chun.chen@wur.nl
Colin Millar	ICES	Denmark	colin.millar@ices.dk
David Currie	Marine Institute	Ireland	david.currie@marine.ie
Duncan Parnell	CEFAS	United Kingdom	duncan.parnell@cefas.co.uk
Edvin Fuglebakk	Institute of Marine Research	Norway	edvin.fuglebakk@hi.no
Hans Gerritsen	Marine Institute	Ireland	hans.gerritsen@marine.ie
Henrik Kjems-Nielsen	ICES	Denmark	henrikkn@ices.dk
Johnathan Ball	CEFAS	United Kingdom	johnathan.ball@cefas.co.uk
Jon Elson	CEFAS	United Kingdom	jon.elson@cefas.co.uk
Josefina Teruel Gómez	Instituto Español de Oceanografía	Spain	josefina.teruel@ieo.es
Julia Wischnewski	Thünen Institute of Sea fisheries	Germany	julia.wischnewski@thuenen.de
Karolina Molla Gazi	Wageningen University & Research	The Netherlands	karolina.mollagazi@wur.nl
Katarzyna Krakówka	National Marine Fisheries Research Institute	Poland	kkrakowka@mir.gdynia.pl
Kirsten Birch Håkansson	Danish Technical University	Denmark	kih@aqua.dtu.dk
Liz Clarke	Marine Scotland	United Kingdom	Liz.Clarke@gov.scot
Marta Suska	National Marine Fisheries Research Institute	Poland	msuska@mir.gdynia.pl
Mary Christman	University of Florida	USA	marycchristman@gmail.com
Nuno Prista	Swedish University Agricultural Sciences	Sweden	nuno.prista@slu.se
Pedro Lino	IPMA	Portugal	plino@ipma.pt
Perttu Rantanen	Natural Resources Institute Finland	Finland	Perttu.Rantanen@luke.fi
Petri Sarvamaa	Natural Resources Institute Finland	Finland	Petri.Sarvamaa@luke.fi
Richard Meitern	University of Tartu	Estonia	richard.meitern@ut.ee
Sven Stoetera	Thünen-Institut	Germany	sven.stoetera@thuenen.de

Annex 2: Agenda

Time (CET)	Monday	Tuesday	Wed	Thursday	Friday
08-09	---	---	---	---	---
09-10	plenary	subgroups	subgroups	subgroups	subgroups
10-11	plenary	subgroups	subgroups	plenary	subgroups
11-12	subgroups	subgroups	subgroups	plenary	subgroups
12-13	*break*	*break*	*break*	*break*	*break*
13-14	subgroups	subgroups	subgroups	subgroups	plenary
14-15	subgroups	subgroups	plenary	subgroups	plenary
15-16	subgroups	subgroups	plenary	subgroups	plenary
16-17	subgroups	subgroups	subgroups	subgroups	---
17-18	subgroups*	subgroups*	subgroups*	social	---
18-19	---	---	---	---	---

colour legend
tor a.
tor b.
tor c.

*chairs to meet with subgroup chairs

Annex 3: Background document for response to special request regarding precision and bias based on RDBES format

Please note: This annex was first published separately as an ad hoc report in December 2020 before the publication of the full WKRDB-EST2 report serving as background documentation for the EU request to ICES on providing output on evaluating data accuracy (precision and bias) for design-based estimation[3].

This report aims to support EU member states in evaluating the statistical accuracy of their catch sampling data, where accuracy refers to the closeness of statistical estimates to their true values. Statistical accuracy is considered in terms of two components: precision and bias. Random uncertainties inherent in estimates due to sampling are described by precision, whilst systematic differences between the estimate and the true value are described by bias. Since this is a complex subject and sampling programmes are usually implemented differently in different countries the work presented relates only to national probabilistic sampling and design-based estimation. To use the code developed, member states will need to convert their national data to the commercial fisheries Regional Database & Estimation System (RDBES) data format.

The evaluation of data precision has been performed using two complementary techniques. For relatively simple sampling designs it is possible to use analytical functions to calculate the precision (or a related statistical measure such as variance) of a statistical estimate. We present these calculations and implementations of these calculations in R code. For more complicated sampling designs, the use of analytical functions is usually not feasible. In these cases, it is necessary to evaluate precision using resampling techniques such as bootstrapping. This report discusses when bootstrapping is appropriate and gives several worked examples describing how bootstrapping can be applied in different cases.

The evaluation of bias in catch sampling is a difficult subject and most biases are generally hard to quantify. It should be noted that there can be several different types of bias occurring at different points in the data collection and the advice production cycle. This report only considers the type of bias that may occur as a result of sampling—not other biases such as those that may be present in particular estimators or stock assessment models. Our approach to bias builds on the previous work available in the ICES literature to identify and enumerate common sources of bias in catch sampling programs. The information was collated and an evaluation performed as to whether data stored using the RDBES data format can inform about potential biases. Reports are presented that can help member states to identify deviations in their sampling programmes and sampling variability that can potentially lead to bias in catch estimates.

The report is a first step towards providing EU member states with a set of tools that can be used to characterize the precision and bias of their catch sampling data. The aim is to provide a solid foundation that, whilst immediately useful in itself, has greater value as a building-block for future work. To this end, a summary of the further activity that is required to extend the work to other scenarios (such as regional sampling programmes) is presented.

---

[3] : ICES. 2020. EU request on providing output on evaluating data accuracy (precision and bias) for design-based estimation at a national level in the form of a report. *In* Report of the ICES Advisory Committee, 2020. ICES Advice 2020, sr.2020.14. https://doi.org/10.17895/ices.advice.7641.

Table of Contents

1 Introduction

The management of the Common Fisheries Policy (CFP) should be guided by the principles of good governance. Those principles include decision-making based on the best available scientific advice, which requires harmonized, reliable, and accurate datasets. To achieve this, EU member states (MS) are obliged to collect and manage data in accordance with the recast Data Collection Framework (DCF) (Regulation (EU) 2017/1004) and the Commission Decisions (EU) 2019/909 and (EU) 2019/910. The DCF places a strong emphasis on cooperation between MS and Regional Coordination Groups (RCGs) are established to support this. Furthermore, MS are encouraged to align their data collection in regional work plans.

It is important to realize that the vast majority of datasets that are used in stock assessment or by other end-users are the result of a series of complex data transformations (e.g. data on unwanted catches). Harmonized, reliable and accurate datasets are thereby not only dependent on data of good quality but also on quality assured processes to transform data. It becomes even more complex in a multinational context where different MS use different processes and where these processes are not always fully documented. The result might be that it is not possible to fully assess the quality of the multinational dataset or the impact the quality has on subsequent analyses. One of the reasons for the present situation is that this processing often requires access to detailed data that might be confidential (e.g. commercial fisheries data).

Harmonized, reliable and accurate datasets also require data collection schemes that are built on sound statistical principles. The MS have worked for several years to establish such schemes in the logistically complex environment that fisheries constitute. The work continues and MS are now also focusing on integrating these new designs into data processing/estimation methods and databases. All this work is a prerequisite for future implementation of regional work plans, as integrated data collection also requires integrated data processing/estimation and management.

1.1 Regional databases supporting the CFP

There is an existing commercial fisheries Regional Database (RDB) that is hosted by ICES and currently used to store aggregated effort and landings data, and detailed biological sampling data. The MS in the North Atlantic, North Sea & Eastern Arctic, and Baltic Sea Regional Coordination Groups currently submit data to the RDB annually and use it to support their work. The Long Distance Fisheries RCG also intends to submit data to the RDB in the near future.

Since the RDB was first developed, the requirements of the DCF have become both broader and more complex. Alongside this, there have been changes in wider fisheries management legislation such as the Landings Obligation. During this time, there have also been a number of improvements within scientific data collection practices including the move towards Statistically Sound Sampling Schemes (“4S”), greater regional coordination, and greater transparency in the scientific evidence base used for fisheries advice. The existing RDB is not able to fully support these new requirements. To this end, a new regional database, the Regional Database & Estimation System (RDBES), has been designed and is in the process of being implemented.

One objective of the RDBES is to support the CFP by improving the harmonisation, transparency and quality assurance of datasets used in analyses underpinning scientific advice. The RDBES is also a prerequisite for the implementation of regional sampling plans and production of transparent regional datasets.

1.2 From the Regional Database (RDB) to the new Regional Database & Estimation System (RDBES)

The Commission is generally supportive of the development of compatible regional databases. This is specified in legislation, especially Article 18 of the recast DCF (Regulation (EU) 2017/1004):

> *"With a view to reducing costs and facilitating access to detailed and aggregated data for end-users of scientific data and other interested parties, Member States, the Commission, scientific advisory bodies and any relevant end-users of scientific data shall cooperate to develop compatible data storage and exchange systems, taking into account the provisions of Directive 2007/2/EC. Those systems shall also facilitate dissemination of information to other interested parties. Such systems may take the form of regional databases. Regional work plans referred to in Article 9(8) of this Regulation may serve as a basis for agreement on such systems."*

The important points about the RDBES development from the Commission's point of view have been enumerated as[4]:

- To ensure regional database functionality for RCG use is uninterrupted;
- That access to data is provided in line with EU policy (MS ownership of data and agreement before use; RCGs have access to the regional database at all times and can use the data; confidentiality rules);
- The Commission supports any extension of the RDBES to other variables (such as recreational fisheries, large pelagics) and other currently separate databases;
- To be able to use the future RDBES for automatic reporting of DCF deliverables such as Annual Reports or Work Plans–National Correspondents should be able to extract data to create the required tables;
- It is important to ensure compatibility between the ICES RDBES and other similar databases (i.e. the proposed Med&BS regional database).

The RDBES is currently in development and is scheduled to go live during 2022, at which point the existing RDB will become read-only.

1.3 Benefits of the RDBES

The aims of the RDBES are:

7. To make data available for the RCGs;
8. Provide a regional estimation system for ICES stock assessments;
9. To increase the data quality, documentation of data, and the use of approved methods;
10. To facilitate the production of fisheries management advice and reports;
11. To increase the awareness of fisheries data collected and the overall usage of these data.

These aims are fully in line with the DCF and support the Common Fisheries Policy (CFP) aim to "*…conserve fish stocks and reduce overfishing in order to provide EU citizens with a long-term stable, secure and healthy food supply.*"[5]

---

[4] ICES. 2020. Steering Committee of the Regional Fisheries Database (SCRDB; outputs from 2019 meeting). ICES Scientific Reports. 2:24. 57 pp. http://doi.org/10.17895/ices.pub.5992

[5] https://ec.europa.eu/info/research-and-innovation/research-area/oceans-and-seas/eu-common-fisheries-policy_en

Aim 1: Make data available for the RCGs

A key aim of the RDBES is to support the DCF work of MS by supporting the RCGs. RCGs are responsible for the coordination of MS sampling activity of commercial fisheries. That sampling is the basis for the estimates of commercial catches used in ICES advice for upcoming years—effective work at RCG level is ultimately needed to fulfil the CFP objectives. The RDBES will allow MS to upload both their commercial detailed biological sample data and aggregated effort and landing data to a new regional database.

The RDBES will support the work of the North Atlantic, North Sea & Eastern Arctic, Baltic Sea, and Long Distance Fisheries RCGs. These RCGs have previously stated their strong support for developing the RDBES. In their 2018 meetings the North Atlantic, North Sea & Eastern Arctic, Baltic Sea RCGs recommended the development and use of the RDBES to store and analyse data[6]. The Long Distance Fisheries RCG also have stated their desire for MS to upload their data to a regional database and require some features of the new RDBES for this to fully occur[7].

The RDBES could potentially also support the work of the recreational, diadromous, and large pelagic regional data collection, but this will be dependent on whether they wish to pursue this and that the funding is available for any developments agreed upon.

As stated earlier Regulation (EU) 2017/1004 encourages the use of compatible regional databases. Both the RDB and the RDBES fulfil this, however, the RDBES has been designed to better allow RCGs and MS to fulfil their obligations towards documenting and improving data quality, and designing and implementing regional sampling designs. The RDBES is designed to capture information about both biological data and how it was sampled. This allows much more realistic analysis of sampling activity to be performed. Similarly to the existing RDB, the RDBES will allow RCGs to analyse data collected by MS at a regional level but new, regional sampling designs will also be supported by the RDBES. The RDBES has been designed to allow the storage of data from regional sampling schemes, in particular by allowing the specification of different sampling designs for different strata within an overall regional sampling design. The estimation system within the RDBES will also allow for the production of regional estimates. The RDBES will also be used by the RCGs to support regional work plans and sampling schemes (such as referred to in Regulation (EU) 2017/1004).

The RDBES will support MS to implement Article 5 of Commission Implementing Decision (EU) 2016/17 by allowing MS to record their statistically sound sampling designs in a common format.

Unlike the RDB, the RDBES will allow sufficient data about the observation of bycatch and Protected, Endangered and Threatened Species (PETS) to be stored such that the relevant bycatch working groups can use it in their work. As Regulation 812/2004 will be repealed, monitored data on bycatch of PETS in commercial fisheries will be included in ICES RDBES. Regulation (EU) 2019/1241 requires the collection of scientific data on incidental catches of sensitive species and the RDBES provides regional storage for this information.

Regulation (EU) 2017/1004 states "It is important to collect biological data on recreational fisheries where there is a potentially significant impact on the state of the stock…" and the RDBES could provide a regional storage system for this recreational data - this regional storage is currently missing because it cannot be stored in the RDB. On the RDBES development, roadmap the

[6] For a summary of the recommendations, see 2018 liaison meeting report: https://datacollection.jrc.ec.europa.eu/documents/10213/1239605/2018-10_15th_Liaison_Meeting.pdf/0d3baf0b-c9a3-410c-936c-a1c7260b6d6d

[7] https://datacollection.jrc.ec.europa.eu/documents/10213/1239599/2019_RCG+LDF.pdf/eb94930a-6fbe-44ac-a833-4b33d63d3e8

possibility of storing recreational data has been included. Whilst it would be good to increase the documentation and transparency of the recreational data used in stock assessments, this is a complicated matter as each of the recreational surveys is conducted in its own way depending on cultural specifications and there is a wide variety of methods being used to sample. The initial plan is therefore to restrict the storage of recreational fisheries data to an aggregated level—with a move to detailed data storage possible in the future.

Aim 2: Provide a regional estimation system for ICES stock assessments

The RDBES will be a key part of ICES stock assessment and the way it supports the CFP objective of conserving fish stocks. ICES stock assessment currently depends on many different data calls and many countries and data submitters, including non-EU countries like Norway and Iceland that jointly contribute national estimates of commercial catches to each ICES stock assessment working group. This means there is a duplication of effort and a lack of consistency. The detailed data and processes used by MS to submit data for stock assessment are also not visible outside of the MS and it is hard for ICES to assure the quality of data provided by sometimes dozens of different individuals for a single fish stock. Important data quality indicators such as estimates of precision are often not submitted to stock assessment data calls. The RDBES will resolve these problems by (i) allowing MS to peer-review each other's estimation algorithms and validate their suitability, (ii) share common functions and tools to obtain those estimates, (iii) strengthening the link between data collectors and stock assessment groups, and (iv) allowing precision of the stock estimates to be correctly calculated and incorporated into the stock assessment models.

Assuring quality is a key element of the ICES advice plan [8]and the RDBES will be an essential tool in the overall quality assurance framework. As a key client of ICES advice and responsible for the success of the CFP, the EU Commission will directly benefit from improvements in ICES stock assessment output.

Aim 3: To increase the data quality, documentation of data, and the use of approved methods

Under Article 14 of Regulation (EU) 2017/1004 MS have a responsibility to ensure that primary, detailed, and aggregated data has appropriate quality assurance and control measures applied before transmission to end-users and that these quality assurance measures have been developed in accordance with the procedures adopted by international scientific bodies, STECF and RCGs.

Generally, each MS has a unique format for its national databases and its own procedures for calculating its estimates of commercial catches so it is very difficult to develop, share, or evaluate data quality tools and estimation methods directly between countries. The RDB was a first step in the right direction and allowed the development of a number of common quality tools and a few standard algorithms of estimation. However, the RDB data format makes strong assumptions about the sampling schemes that MS are using that frequently differ from the way national data is actually collected. This has made it difficult for data to be analysed within the RDB. Unlike the current RDB, the new RDBES stores all the important information about how data was collected including all the novel statistically sound sampling variables demanded in Commission Implementing Decision (EU) 2016/1701. These variables include, but are not restricted to, the sampling scheme used, the sampling frames and stratification schemes, the different sampling units (e.g. fishing trips, port-days), and how units were selected for sampling (e.g. simple random sampling, expert judgement) in all sampling levels. This will allow new and better estimates of commercial catches and accompanying quantitative measures of quality to be calculated using

[8] https://issuu.com/icesdk/docs/ices_advisory_plan

the RDBES (e.g. the precision of any estimates calculated from the data). It also allows documentation about the different sampling schemes that MS are using to be easily generated.

Using a common regional platform to develop quality checks and having a common, transparent and documented repository of estimation algorithms used in producing the commercial data entering stock assessment also means that MS can develop these procedures in a collaborative peer-reviewed manner which will improve efficiency and effectiveness. For example, it will be possible to encode approved statistical quality checks and estimation procedures that many other MS can review and use. Data quality checks that can be applied consistently on a regional scale will also be necessary for the regional work programmes that are currently being developed.

Aim 4: To facilitate the production of fisheries management advice and reports

The RDBES web application will provide certain functionality such as data uploading, and managing permissions but stock estimation and imputation will be performed within the ICES Transparent Assessment Framework (TAF) - this is an open framework for organising stock assessments. All data inputs and outputs are traceable and versioned. The open framework enables stock assessment scientists to easily find, reference, download, and run the assessment from any stage in the process leading to the published ICES advice for a given stock. Anyone will also be able to find, reference, and download the estimation method behind the assessment (but not the underlying data). Basing the stock estimation functions of RDBES on the TAF has a number of advantages: the TAF exists and users are already gaining expertise in it, there is technical and content support available, version control of data and scripts is established, and it provides strong linkages to stock assessment groups.

A key benefit of the RDBES is that it will be used to assure the quality of the DCF process from data collection to stock assessment.

Aim 5: To increase the awareness of fisheries data collected and the overall usage of these data

> *"[Regional databases]…facilitate the work of the EU Member States by reducing the burden of multiple data submissions (for data calls) under different formats. They allow end users to calculate statistical estimates of data tailored to their needs, and help to streamline and ease the reporting of Member States on the EU data collection."*[9]

The aim of the RDBES is that data are available at the highest possible resolution whilst taking into account data ownership, access rights, and confidentiality constraints. This means that it could be possible to use the data for other relevant purposes. For example, currently, MS submit data to both the RDB data call and the Fisheries Dependent Information (FDI) data calls. This involves both a duplication of effort and can produce a lack of consistency. Unlike the existing RDB, the aggregated effort and landings data model in the new RDBES has been designed to be compatible with the FDI data call. Thus it could be possible for MS to use the RDBES to also respond to the FDI data call. In the same way, it should also be possible for MS to use the RDBES to complete part of their DCF annual reports. This would be a benefit for MS since they would not have to duplicate all the data submission work, and a benefit for the STECF since it would remove a possible source of consistency error.

[9] Annex 3 *in* "Call for proposals MARE/2020/08 Strengthening regional cooperation in the area of fisheries data collection" http://ec.europa.eu/newsroom/document.cfm?doc_id=66541

1.4 Development of RDBES tools to evaluate data accuracy relating to bias and precision

As previously discussed, the annual national work plans and reports of MS are an important record of the data quality processes that are applied at the national level. Specifically, this information is summarized in Table 5A of the EU-MAP. This table typically asks whether documentation on a subject exists and, if so, where that documentation can be found. The subjects covered include sampling design, quality checks at the point of data capture, evaluation of precision and bias, and editing and imputation methods. The contents of these tables have been analysed during inter-sessional work of the RCGs and it has been seen that MS have difficulty answering some of these questions since there is a lack of guidance or tools available on the subject.

In particular, the documentation around data accuracy, bias and precision has been observed to be one of the weaker areas - specifically related to the following questions[10]:

- "Are processes to evaluate data accuracy (bias and precision) documented?"
- "Where can documentation on processes to evaluate accuracy be found?"

When completing this table one MS commented,

> *"Presently, we do not evaluate bias and precision of our data because we are not aware of routine tools available for such estimates on a national level. As soon as routines are available we will use these. (...)"*

It can be seen that tools to evaluate data accuracy relating to bias and precision at a national level are required. Evaluation of this bias and precision at a national level will also be relevant to ICES and the Commission since these data feed into stock assessments and can affect the accuracy of their outputs. To enable this advice to be used by all MS (and ICES member countries if they desire) it should be based on a common data format from which statistical bias and precision can be correctly calculated. The new RDBES data model provides that format since it provides a common structure to describe both the detailed sampling data and, importantly, the sampling design underlying how those data were obtained.

Still, just having data in a sophisticated data structure like the RDBES is not enough: the very estimation of precision and bias for individual programmes is a complex subject frequently found diversely implemented in different countries. For example, there are a number of different estimation techniques that can be used to create inputs for stock assessment from biological data. Broadly these can be categorized as "model-based" and "design-based" estimation methods. (Model-based methods are in common use but involve assumptions on sampling as well as on nature which can be difficult to verify whereas design-based estimators involve only assumptions on sampling which are in principle controllable and easier to scrutinize.)

To resolve this, in the first instance the tools will relate specifically to design-based estimation since substantial further work will be required for it to be applied to other types of estimation. A roadmap has been produced for the work required to extend the tools to these other types of estimation in the future.

This report on evaluating data accuracy (precision and bias) for design-based estimation at a national level covers the following subjects:

---

[10] RCG NA NS&EA RCG Baltic 2020. Regional Coordination Group North Atlantic, North Sea & Eastern Arctic and Regional Coordination Group Baltic. 2020. Part I Report, 110 pgs. Part II Decisions and Recommendations, 7 pgs. Part III, Intersessional Subgroup (ISSG) 2019-2020Reports, 154 pgs. See: https://datacollection.jrc.ec.europa.eu/docs/rcg

- Definition of the prerequisites that a MS will need to meet to be able to use the tools (e.g. MS data will need to be in the RDBES data format; the MS will need to be carrying out probabilistic sampling and recording certain data);
- Specification of the statistical functions to allow MS to evaluate bias and estimate precision for design-based estimation;
- Identification of further functions that would be required in the future to evaluate data accuracy for other type of estimation, and for regional data estimation;
- Recommendations for further work and a roadmap of how to extend the advice to other types of bias and precision estimation.

1.5 Summary

This section has shown how the new RDBES that is currently in development will be better able to support the recast EU Data Collection Framework (Regulation (EU) 2017/1004) than the existing RDB. The RDBES will provide an essential platform for MS and RCGs to fulfil their obligations towards documenting and improving data quality and designing and implementing regional sampling designs. ICES is an important end-user of DCF data and the RDBES will be a key input to ICES stock assessments. In particular, the RDBES will improve data quality and transparency by allowing peer-review of procedures, sharing common functions and tools to obtain those estimates, and allowing precision of stock estimates to be correctly calculated and incorporated into the stock assessment models. As a key client of ICES advice and responsible for the success of the CFP, the EU Commission will directly benefit from improvements in ICES stock assessment output.

2 Data accuracy

2.1 Introduction

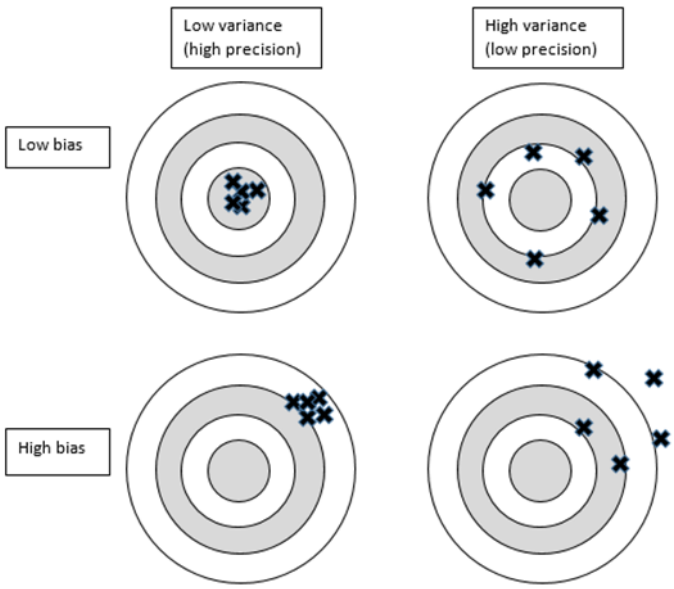
This section describes how the general concept of data accuracy is treated within statistical analyses and defines the scope of the types of fisheries sampling that this report covers, and the approach used.

It is useful to first define what is meant by data accuracy in this context:

> *"Accuracy of data is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure...The concept of accuracy relates a numerical estimate to its true value according to an agreed definition. The closer the estimate is to its true value, the more accurate it is. The difference between the estimate and the true value is called the error of the estimate and error is thus a technical term to represent the degree of lack of accuracy. The error has a random component (variance) as well as a systematic component (bias). It is sometimes better to speak of uncertainty than error, when the term error risks to be confused with a mistake committed, which is a very different matter."*(European Statistical System (ESS) handbook for quality and metadata report, 2020, p.98)

In the context of Table 5A within the DCF National Workplans / Annual Reports the concept of data accuracy is explicitly linked with the terms "precision" and "bias". In this case, precision can be considered to be inversely related to variance i.e. a higher variance in the random component of the uncertainty means a lower precision.

An informal example, which is often given to illustrate the difference between variance and bias, is that of trying to shoot arrows at a target. Ideally, we would like all our arrows to be in the centre. The diagram below illustrates how the arrows might hit the target in different variance and bias scenarios:



Clearly, the desired situation is to have both low variance (high precision) and low bias in our estimates although this may not always be possible in practice.

It should be noted that there can be a number of different types of bias occurring at different points in the data collection and advice production cycle – in this report we only consider bias that may occur as a result of sampling, not other biases such as those that may be present in particular estimators, or stock assessment models.

2.2 Scope

The aim of this work is to produce a first step towards creating general tools that MS can use to evaluate data accuracy but it is not possible to cover all scenarios given the time and resources available. It is thus necessary to restrict the applicability of this work to the following requirements:

12. The data is collected by commercial fisheries sampling programmes performed by a single institute;
13. The sampling programmes considered should be probabilistic;
14. The sampling can be multi-stage, with stratification at any or all levels. Units may be selected with or without replacement;
15. Estimation of the desired parameters should be by design-based estimation;
16. The data is available in the ICES Regional Database & Estimation System (RDBES) format[11];
17. The sampling data should not require the use of any of the cluster sampling variables defined within the RDBES data model[12].

2.3 Approach

The evaluation of data precision has been performed using two complementary techniques. For relatively simple sampling designs it is possible to use analytical functions to calculate the precision (or a related statistical measure such as variance) of a statistical estimate. In Section 3 we present these calculations and implementations of them in R code. For more complicated sampling designs, the use of analytical functions is usually not feasible. In these cases, it is necessary to evaluate precision using resampling techniques such as bootstrapping. Section 4 discusses when bootstrapping is appropriate and gives a number of worked examples describing how bootstrapping can be applied in different cases.

The evaluation of bias is a difficult subject and is hard to quantify. The approach presented in Section 5 builds on the previous work available in the ICES literature to identify and enumerate common sources of bias in catch sampling programs. The information is collated and an evaluation is then performed as to whether data stored using the RDBES data format can inform about that bias source. Reports are presented that can help member states to identify deviations in their sampling programmes and sampling variability that can potentially lead to bias in catch estimates.

[11] https://github.com/ices-tools-dev/RDBES

[12] XXselectionMethodCluster, XXnumberTotalClusters, XXnumberSampledClusters, XXselectionProbCluster, XXinclusionProbCluster

3 Analytical calculation of variance

This section presents the variance calculations for design-based estimation using a three-stage sampling design.

3.1 Sampling without replacement in all three stages

Consider the following sampling design in three stages where the primary sampling units are vessels, the secondary sampling units are trips and the tertiary sampling units are hauls.

Stage I: Sampling of vessels
A random sample without replacement of vessels is drawn from all the vessels in the population. The set of vessels in the population is denoted U_I of size N_I and the sample of vessels is denoted s_I of size n_I. Each vessel is looked upon as a cluster of trips.

Stage II: Sampling of trips
For every vessel i selected in stage I, a random sample without replacement of trips is drawn from all the trips associated with the vessel. The set of trips associated with vessel i is denoted U_{IIi} of size N_{IIi} and the sample of trips is denoted s_{IIi} of size n_{IIi}. Each trip is looked upon as a cluster of hauls.

Stage III: Sampling of hauls
For every trip q selected in stage II, a random sample without replacement of hauls is drawn from all the hauls associated with the trip. The set of hauls associated with trip q is denoted U_{iq} of size N_{iq} and the sample of hauls is denoted s_{iq} of size n_{iq}.

For each haul k selected in stage III, the weight of discards, y_k, is observed. The problem is to estimate the total weight of discards for all possible hauls, trips and vessels,

$$t_y = \sum_{U_I} \sum_{U_{IIi}} \sum_{U_{iq}} y_k$$

and the variance of this estimator. To accomplish this, we need the inclusion probabilities for each stage.

3.1.1 Inclusion probabilities for the general case

For **stage I**, the first order inclusion probability π_{Ii} is the probability of vessel i to be included in the sample s_I. The second-order inclusion probability π_{Iij} is the joint probability of vessel i and j to be included in s_I.

For **stage II**, the first order inclusion probability $\pi_{IIq|i}$ is the conditional probability of trip q to be included in the sample s_{IIi} (conditional on the stage I sampling). The second-order inclusion probability $\pi_{IIqr|i}$ is the conditional joint probability of trip q and r to be included in s_{IIi}.

For **stage III**, the first order inclusion probability $\pi_{k|iq}$ is the conditional probability of haul k to be included in the sample s_{iq} (conditional on stage I and II sampling). The second-order inclusion probability $\pi_{kl|iq}$ is the conditional joint probability of haul k and l to be included in s_{iq}.

We summarize these general inclusion probabilities in the table below:

Inclusion probabilities, general		
Stage	First-order	Second-order
I	π_{Ii}	π_{Iij}
II	$\pi_{IIq\|i}$	$\pi_{IIqr\|i}$
III	$\pi_{k\|iq}$	$\pi_{kl\|iq}$

(Note that $\pi_{Iii} = \pi_{Ii};\ \pi_{IIqq|i} = \pi_{IIq|i};\ \pi_{kk|iq} = \pi_{k|iq}$.)

3.1.2 Estimation for the general case

In general, the Horvitz-Thompson (HT) estimator of t_y with respect to all three stages is given by

$$\hat{t}_y = \sum_{s_I} \frac{1}{\pi_{Ii}} \sum_{s_{IIi}} \frac{1}{\pi_{IIq|i}} \sum_{s_{iq}} \frac{y_k}{\pi_{k|iq}}$$

We can also write $\hat{t}_y$ as

$$\hat{t}_y = \sum_{s_I} \frac{\hat{t}_i}{\pi_{Ii}}$$

where $\hat{t}_i$ is the HT estimator of the total weight of discards for vessel i with respect to stage II and III:

$$\hat{t}_i = \sum_{s_{IIi}} \frac{1}{\pi_{IIq|i}} \sum_{s_{iq}} \frac{y_k}{\pi_{k|iq}}$$

Similarly, the estimator $\hat{t}_i$ can be written as

$$\hat{t}_i = \sum_{s_{IIi}} \frac{\hat{t}_{iq}}{\pi_{IIq|i}}$$

where $\hat{t}_{iq}$ is the HT estimator of the total weight of discards for trip q with respect to stage III:

$$\hat{t}_{iq} = \sum_{s_{iq}} \frac{y_k}{\pi_{k|iq}}$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = \sum\sum_{s_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}$$

where

$$\hat{V}_i = \sum\sum_{s_{iq}} \frac{\pi_{kl|iq} - \pi_{k|iq}\pi_{l|iq}}{\pi_{kl|iq}} \frac{y_k}{\pi_{k|iq}} \frac{y_l}{\pi_{l|iq}}$$

Note that for the point estimator we only use the first order inclusion probabilities. For the variance estimator, we also need the second-order inclusion probabilities.

3.1.3 Inclusion probabilities for Simple Random Sampling (SRS) without replacement in each stage

The inclusion probabilities valid for the case of SRS without replacement in each stage are given in the table below.

	Inclusion probabilities, SRS without replacement	
Stage	**First-order**	**Second-order**
I	$\frac{n_I}{N_I}$	$\frac{n_I(n_I-1)}{N_I(N_I-1)}$
II	$\frac{n_{IIi}}{N_{IIi}}$	$\frac{n_{IIi}(n_{IIi}-1)}{N_{IIi}(N_{IIi}-1)}$
III	$\frac{n_{iq}}{N_{iq}}$	$\frac{n_{iq}(n_{iq}-1)}{N_{iq}(N_{iq}-1)}$

3.1.4 Estimation for SRS without replacement in each stage

For SRS without replacement in each stage, the HT estimator of t_y simplifies into

$$\hat{t}_y = \sum_{s_I} \frac{N_I}{n_I} \sum_{s_{IIi}} \frac{N_{IIi}}{n_{IIi}} \sum_{s_{iq}} \frac{N_{iq}}{n_{iq}} y_k = \frac{N_I}{n_I} \sum_{s_I} \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} \frac{N_{iq}}{n_{iq}} \sum_{s_{iq}} y_k$$

The estimator can also be written as

$$\hat{t}_y = \frac{N_I}{n_I} \sum_{s_I} \hat{t}_i$$

where

$$\hat{t}_i = \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} \frac{N_{iq}}{n_{iq}} \sum_{s_{iq}} y_k = \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} \hat{t}_{iq}$$

and

$$\hat{t}_{iq} = \frac{N_{iq}}{n_{iq}} \sum_{s_{iq}} y_k$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = N_I^2 \frac{1 - n_I/N_I}{n_I} S^2_{\hat{t}s_I} + \frac{N_I}{n_I} \sum_{s_I} \left[N_{IIi}^2 \frac{1 - n_{IIi}/N_{IIi}}{n_{IIi}} S^2_{\hat{t}s_{IIi}} + \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} N_{iq}^2 \frac{1 - n_{iq}/N_{iq}}{n_{iq}} S^2_{ys_{iq}} \right]$$

where

$$S^2_{\hat{t}s_I} = \frac{1}{n_I - 1} \sum_{s_I} \left[\hat{t}_i - \left(\sum_{s_I} \hat{t}_i / n_I \right) \right]^2;$$

$$S^2_{\hat{t}s_{IIi}} = \frac{1}{n_{IIi} - 1} \sum_{s_{IIi}} \left[\hat{t}_{iq} - \left(\sum_{s_{IIi}} \hat{t}_{iq} / n_{IIi} \right) \right]^2;$$

$$S^2_{ys_{iq}} = \frac{1}{n_{iq} - 1} \sum_{s_{iq}} \left[y_k - \left(\sum_{s_{iq}} y_k / n_{iq} \right) \right]^2$$

3.1.5 Simplified variance estimation

Some simplified variance estimators for multistage sampling are discussed in Särndal *et al*. (1992, section 4.6). One possibility is to use only the first term in the expression for the variance estimator; that is, using the abridged HT estimator

$$\hat{V}(\hat{t}_y) = \sum\sum_{s_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Iij}}{\pi_{Iij}} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}}$$

Under SRS without replacement in all stages, this would mean using

$$\hat{V}(\hat{t}_y) = N_I^2 \frac{1 - n_I/N_I}{n_I} S^2_{\hat{t}s_I}$$

If the sample size in stage I is fixed, an alternative is to use the abridged Yates-Grundy estimator

$$\hat{V}^*(\hat{t}_y) = -\frac{1}{2}\sum\sum_{s_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Iij}}{\pi_{Iij}} \left(\frac{\hat{t}_i}{\pi_{Ii}} - \frac{\hat{t}_j}{\pi_{Ij}}\right)^2$$

In both cases, this would lead to underestimation of the true variance. However, if the variance contributions from stage II and III are small, this underestimation might not be so important.

Another option is to do the variance estimation as if vessels were selected with replacement in stage I. The estimation formula for this situation is given in the next section. This approach might in general lead to both over- and underestimation of the true variance.

3.2 Sampling with replacement in the first stage

Consider again a sampling design in three stages where the primary sampling units are vessels, the secondary sampling units are trips and the tertiary sampling units are hauls. The difference from the design in section 1 is that the sampling is done with replacement in the first stage whereas the sampling in subsequent stages is still without replacement.

Stage I: Sampling of vessels

A random sample *with replacement* of vessels is drawn from all the vessels in the population in such a way that, at every draw, p_i is the probability of selecting vessel i. The set of vessels in the population is denoted U_I of size N_I. The ordered sample of vessels is denoted $os_I = (i_1, \dots, i_v, \dots, i_{m_I})$, where i_v is the vessel selected in draw number v and m_I is the number of draws. Each vessel is looked upon as a cluster of trips.

Stage II: Sampling of trips

For every vessel drawing i_v in stage I, a random sample without replacement of trips is drawn from all the trips associated with the vessel. The set of trips associated with vessel drawing i_v is denoted U_{IIi_v} of size N_{IIi_v} and the sample of trips is denoted s_{IIi_v} of size n_{IIi_v}.

Stage III: Sampling of hauls

For every trip q selected in stage II, a random sample without replacement of hauls is drawn from all the hauls associated with the trip. The set of hauls associated with trip $q \in s_{IIi_v}$ is denoted U_{i_vq} of size N_{i_vq} and the sample of hauls is denoted s_{i_vq} of size n_{i_vq}.

We assume that the sampling in stage II and III has the properties of invariance and independency.

3.2.1 Estimation for the general case

In general, the Hansen-Hurwitz (HH) estimator of t_y with respect to all three stages is given by

$$\hat{t}_y = \frac{1}{m_I}\sum_{v=1}^{m_I} \frac{\hat{t}_{i_v}}{p_{i_v}}$$

where $\hat{t}_{i_v}$ is the HT estimator of the total weight of discards for vessel drawing i_v with respect to stage II and III:

$$\hat{t}_{i_v} = \sum\nolimits_{s_{IIi_v}} \frac{1}{\pi_{IIq|i_v}} \sum\nolimits_{s_{i_vq}} \frac{y_k}{\pi_{k|i_vq}}$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = \frac{1}{m_I(m_I - 1)} \sum_{v=1}^{m_I} \left(\frac{\hat{t}_{i_v}}{p_{i_v}} - \hat{t}_y \right)^2$$

(see: Särndal *et al.*, 1992, Result 4.5.1).

3.2.2 Estimation for SRS with replacement in the first stage

For SRS with replacement in stage I, the drawing probability p_{i_v} is equal to $1/N_I$ for all vessel drawings i_v. If SRS without replacement is used in stage II and III, the HH estimator of t_y simplifies into

$$\hat{t}_y = \frac{1}{m_I} \sum_{v=1}^{m_I} \frac{\hat{t}_{i_v}}{p_{i_v}} = \frac{N_I}{m_I} \sum_{v=1}^{m_I} \hat{t}_{i_v}$$

where

$$\hat{t}_{i_v} = \frac{N_{IIi_v}}{n_{IIi_v}} \sum\nolimits_{s_{IIi_v}} \frac{N_{i_vq}}{n_{i_vq}} \sum\nolimits_{s_{i_vq}} y_k$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = \frac{N_I^2}{m_I(m_I - 1)} \sum_{v=1}^{m_I} (\hat{t}_{i_v} - \hat{\bar{y}}_U)^2$$

where $\hat{\bar{y}}_U = \hat{t}_y / N_I$.

3.2.3 References

Särndal, C.-E., Swensson, B., Wretman, J. (1992) Model Assisted Survey Sampling. Springer-Verlag.

3.3 Implementation in R

Annex A3.2 below presents an implementation of these analytical variance calculations using the RDBES data model. It specifically considers Hierarchy 1 but is generalizable to all hierarchies.

4 Bootstrapping

4.1 Introduction

Bootstrapping (Efron and Tibshirani, 1986) in the context of catch sampling is a simulation method for approximating the sampling design and estimation procedures commonly used for providing stock assessors with desired means and variances. In the case of the EU, the designs usually involve multistage sampling, often stratified at one or more stages of sampling, and the estimation procedures are usually design-based.

The general approach when bootstrapping is to,

18. Use the original sample data to obtain the estimates of interest;
19. Repeatedly sample from the original dataset, each time following the original sampling design and estimation procedures as much as possible;
20. Using those "bootstrapped" estimates from (2) to assess bias or to estimate sampling variability of the estimates calculated in step (1).

The basic idea behind bootstrapping is that the original sample (S) is representative of the population (P) from which it was taken, where "representative" means that the sample is sufficiently large to capture the range of values in P, the variability inherent in P, and the frequency distribution of those values in P. Hence, S acts as a pseudo-population from which one can repeatedly sample to observe the behaviour of the design and estimators. The approach works well when the sample sizes are large and the sample data cover the characteristics of the population from which they were taken.

On the other hand, S can be a possibly poor representation of P in several situations, ranging from small sample sizes to poorly chosen strata with mismatched strata sample sizes to poor selection procedures, such as always choosing the fishing operations with the largest landings because one wishes to obtain as many species as possible. The issue of deciding whether a sample size is appropriate or sufficiently large is a difficult one as it depends on 1) the size of the sampling frame, e.g. 10 hauls/trip and 3 hauls are sampled, 2) the choices for stratum sample sizes, e.g. one stratum has a large sampling frame and another a small one but the same number of units were sampled in each, 3) the variability of the sampling frame. For example, the sample of 3 hauls on a fishing trip with 10 hauls can be bootstrapped if it is reasonable to assume that the unsampled trips have similar characteristics and variability displayed in the 3 sampled trips. It would not be appropriate if the 3 sampled hauls were always chosen because of some characteristic that distinguished them from the other hauls on the trip. One must keep in mind that even if the analyst is comfortable with the sample sizes used in the bootstrapping, the entire exercise is based on the observed sample. If that sample is not fully representative, then one would not obtain similar results if a different sample had been used. This is of course true even if the estimates of interest (e.g. standard errors) were analytically calculated since the entire exercise is based on the data available.

In addition to the issue of representation, one cannot estimate variance among sampling units using a design-based estimation approach when only one sampling unit is chosen. Hence, like the analytical method, the variance is underestimated due to the lack of information for one or more stages of the design.

Sampling may not be fully appropriate for bootstrapping when the population is large but a very small sample was taken, e.g. when P contains 1000 vessel × trips but only 8 were sampled. It is unlikely that one can argue successfully that the 8 observations are representative of the entire

population of vessel × trips. Again, this is true whether the analyst is bootstrapping or using an analytical method; the analyst needs to be aware that the bootstrapping cannot overcome the effect the small sample size has on the resulting estimates.

Another instance where bootstrapping will not correct internal flaws in the dataset is when the sampling design cannot be replicated using computer code. An example of this is convenience sampling. Hence, bootstrapping is unlikely to be of use for any non-probabilistic method of obtaining S that requires strict assumptions (e.g. assuming the convenience sample is a reasonable facsimile of SRSWR) unlikely to be true.

Bootstrapping works best when the sample sizes are sufficiently large that one is comfortable that S is representative of P, when the sample data were collected according to a design strategy that can be replicated in computer code, and when the estimation procedures can be reproduced without error.

In the case of probabilistic (simple random or unequal probability) with replacement sampling at all stages of the design, the original sample dataset is the pseudo-population. Multistage sampling can be done directly and repeatedly on the original sample following the original design and original sample sizes at each stage and in each stratum. For each new "bootstrap" multistage sample dataset, the desired estimates can be calculated and stored. After a large number of repetitions of the sampling from the pseudo-population, the bias and precision of the estimators used on the original data can be assessed by calculating the mean and variance of the bootstrapped estimates. Pseudo-code for performing this bootstrapping procedure using the sampling design in Example 1 of Hierarchy 1 in the RDBES Data Model documentation is given in the next section and some R code (Box 2) for a slightly different multi-stage sampling design is shown in Annex A3.3 below.

In the case of probabilistic (simple random or unequal probability) where sampling is without replacement at one or more stages of the design, the original sample dataset is used to construct a new pseudo-population. The pseudo-population for the stage sampled without replacement contains multiple copies of the original dataset S from that level. We need a bit of notation here. Let the population have N sampling units and an SRSWOR of n vessels is selected. In the simplest case, suppose $\frac{N}{n} = k$ is an integer (e.g. $N = 100, n = 20$, then $k = 5$). Then, the pseudo-population is simply the set of k replications of the sample. So, in the case of SRSWOR of PSUs but SRSWR at all lower stages, the pseudo-population would be the k copies of the sample of PSUs and all child sample data associated with those PSUs.

In the case where sampling is SRS but k is not an integer (e.g. $N = 100, n = 15$, then $k = 6.667$), there are three approaches possible. First, if the sample size is small relative to the population size, then the SRSWOR is treated as an SRSWR and the pseudo-population is the sample. A commonly used cutoff is to treat an SRSWOR as being with replacement when $\left(100 \times \frac{n}{N}\right)\% < 5$. Second, is to construct a pseudo-population that contains $\lfloor k \rfloor$ copies of S ($\lfloor a \rfloor$ is the largest integer less than a) plus $N - n \times \lfloor k \rfloor$ randomly chosen units from S. In this approach, the pseudo-population has two sources of variability: the original sampling variability plus a variability due to the selection of the additional units to fill out the population to its full size. Hence, any bootstrapping based on this method usually includes a loop to repeatedly create new pseudo-populations and to perform bootstrapping on each. We do not recommend this approach because the pseudo-population characteristics such as its mean and variance vary among the different random realizations of the pseudo-population. The third method is a simpler and more useful approach to creating a pseudo-population from which bootstrap SRSWOR can be taken. In this approach (Bickel and Freedman, 1984; Chao and Lo, 1985; Sitter, 1992) there is randomization between two different pseudo-populations made up of either $\lfloor k \rfloor$ or $(\lfloor k \rfloor + 1)$ copies of the sample S so that in either case, the mean of each of the pseudo-populations matches the mean of S.

We recommend the Sitter (1992) method for constructing the pseudo-population since it has some nice behaviours relative to the bootstrapping that is planned. Some example R code is provided in Boxes 3 and 4 in Annex A3.3 below to demonstrate the method for single-stage and multi-stage designs.

When constructing pseudo-populations for multi-stage sampling designs where SRSWOR occurs at more than one level, the recommended approach (Sitter, 1992) is to construct the pseudo-population for each stage with WOR sampling before and during bootstrapping. For example, suppose both the PSUs and SSUs are selected using SRSWOR and the TSUs are selected by SRSWR. The PSUs might be vessel × trips, the SSUs, fishing operations nested within each vessel × trip, and the TSUs are individual fish (or possibly samples of X buckets of unsorted fish) that are treated as having been selected by SRSWR. Construction of the pseudo-population would start by replicating the sampled SSUs the required number of times to "fill out" all operations for each sampled PSU following the Sitter procedure. Once each sampled PSU has been fully recreated and the desired quantities obtained for that SSU, the pseudo-population of PSU would be constructed by again following the method described by Sitter (1992) and then re-sampling from the pseudo-population.

Once the pseudo-population is created for a multi-stage sampling design, bootstrapping proceeds similar to that for SRSWR. Multistage sampling is done directly and repeatedly on the pseudo-population following the original design and appropriate sample sizes at each stage and in each stratum. For each new "bootstrap" multistage sample dataset, the desired estimates can be calculated and stored. After many repetitions from the pseudo-population, the bias and precision of the estimators used on the original data can be assessed by calculating the mean and variance of the bootstrapped estimates. R code for performing the SRSWOR bootstrapping procedure using the sampling design in Example 1 of Hierarchy 1 is given in Box 4 in Annex A3.3 below.

4.2 Types of data needed for input to bootstrapping catch sampling

For the examples given in this document, there are generally two or three datasets needed for input to the bootstrapping simulations. Tables 4.1–4.3 list the more common variables needed for calculating numbers at length, numbers at age and total discard weight for a given species within a given fishery within a stratum. The tables assume the stratum is quarter; modifications can be made if there are other strata at different stages of the sampling. The lists assume that fish are sampled for length from a fishing operation from a trip on a vessel within a stratum.

Table 4.1. Variables needed to estimate total discards by stratum (quarter in the pseudo-code example). The list is based on providing an estimate for a single year.

Variable Name	Definition
Vessel ID	Unique identifier of the sampled vessel from the reference fleet
Quarter	1, 2, 3, 4
Trip ID	Unique identifier of the sampled trips from the vessel in the quarter
Fishing Operation ID	Unique identifier of the sampled FO within the sampled trip
Species ID	Unique identifier of the species for which total discard weight is of interest
Discard Weight	Observed discard weight for the species in the sampled FO

Variable Name	Definition
Landed Weight	Observed landed weight for the species in the sampled FO
Total Landed Weight	Reported total landed weight for the entire fishery from which vessels, trips and FOs were sampled in the listed quarter

Table 4.2. Variables needed to estimate numbers at length (NAL) by stratum. The list is based on providing an estimate for a single year. Additional information could be required if there is a need for one or more conversions (such as converting fork length to total length).

Variable Name	Definition
Vessel ID	Unique identifier of the sampled vessel from the reference fleet
Quarter	1, 2, 3, 4
Trip ID	Unique identifier of the sampled trips from the vessel in the quarter
Fishing Operation ID	Unique identifier of the sampled FO within the sampled trip
Catch Category	Catch category (DIS, LAN, BMS, ALL, …..)
Species ID	Unique identifier of the species for which total discard weight is of interest
Length Class	Number indicating an observed length class for the species in this FO
Length Unit	mm, cm, scm, ….
Length Type	Type of measurement (total, fork, …)
Number at Length	Number of fish in the length class in the FO sample
Sample Weight of FO	Weight of the sample in the listed catch category in the FO
Total Weight of FO	Total weight of the species in the listed catch category in the FO
Total Weight of Trip	Total weight of the species in the listed catch category in the sampled trip over all FOs, sampled or not
Total Weight of Stratum	Total weight of the species in the listed catch category over all trips, sampled or not, within the quarter

Table 4.3. Variables needed to estimate age-length-key (ALK) by stratum. The list is based on providing an estimate for a single year and assumes that numbers at length are separately calculated before creating the numbers at age (NAA) for the stratum.

Variable Name	Definition
Vessel ID	Unique identifier of the sampled vessel from the reference fleet
Quarter	1, 2, 3, 4
Trip ID	Unique identifier of the sampled trips from the vessel in the quarter
Fishing Operation ID	Unique identifier of the sampled FO within the sampled trip
Catch Category	Catch category (DIS, LAN, BMS, ALL, …..)

Variable Name	Definition
Species ID	Unique identifier of the species for which total discard weight is of interest
Fish ID	Unique identifier for an individual fish in the sampled species in this sampled OF in the sampled trip
Length Class	Number indicating an observed length class for the species in this FO
Length Unit	mm, cm, scm, ….
Length Type	Type of measurement (total, fork, …)
Age Class	Number indicating the estimated age of the fish sampled in the FO on this trip
Age Unit	Year, month, …..

4.3 Notation

$x(y)$ indicates x is nested within y, with Area/Quarter Strata

(note that simpler stratification would remove some of the subscripts)

N_a = total number of vessels in reference fleet in the area stratum $a, a = 1, \dots, A$ (A is the total number of areal reference fleets)

n_a = number of vessels selected from the reference fleet in the area stratum a

$T_{v(a,q)}$ = total number of trips taken by vessel v within area a, $v(a,q) = 1,2, \dots, n_a$, and quarter $q, q = 1,2,3,4$ (note that this could be 0 for some combinations of quarters, areas and vessels)

$t_{v(a,q)}$ = number of trips observed for vessel v within area a and quarter q

$H_{t(v,a,q)}$ = total number of fishing operations performed on trip t for vessel v within area a and quarter q

$LW_{h(t,v,a,q)}$ = total landings weight for the h^{th} fishing operation performed on trip t for vessel v within area a and quarter q (reported), $h = 1, \dots, H_{t(v,a,q)}$

$DW_{h(t,v,a,q)}$ = total discard weight for the h^{th} fishing operation performed on trip t for vessel v within area a and quarter q (usually estimated), $h = 1, \dots, H_{t(v,a,q)}$

$h_{t(v,a,q)}$ = number of fishing operations observed on trip t for vessel v within area a and quarter q

$Lw_{h(t,v,a,q)}$ = sampled landings weight for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q (reported), $h = 1, \dots, h_{t(v,a,q)}$

$Dw_{h(t,v,a,q)}$ = sampled discard weight for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q (usually estimated), $h = 1, \dots, h_{t(v,a,q)}$

$FL_{h(t,v,a,q)}$ = total number of fish in the sampled landings for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q (usually estimated), $h = 1, \dots, h_{t(v,a,q)}$

$fL_{h(t,v,a,q)}$ = number of fish sampled for length in the sampled landings for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q, $h = 1, \dots, h_{t(v,a,q)}$

$FD_{h(t,v,a,q)}$ = total number of fish in the sampled discards for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q (usually estimated), $h = 1, \dots, h_{t(v,a,q)}$

$fD_{h(t,v,a,q)}$ = number of fish sampled for length in the sampled discards for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q, $h = 1, \ldots, h_{t(v,a,q)}$

$faL_{l,h(t,v,a,q)}$ = number of fish sampled for age within length class l, $l = 1, \ldots, \min(X, fL_{h(t,v,a,q)})$, (where X is a species-specific upper limit of the number of fish ideally sampled from each length class) in the sampled landings for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q, $h = 1, \ldots, h_{t(v,a,q)}$

$faD_{l,h(t,v,a,q)}$ = number of fish sampled for age within length class l, $l = 1, \ldots, \min(X, fD_{h(t,v,a,q)})$, (where X is a species-specific number of fish to be sampled from each length class) in the sampled discards for the h^{th} fishing operation sampled on trip t for vessel v within area a and quarter q, $h = 1, \ldots, h_{t(v,a,q)}$

4.4 Pseudo-code for running bootstrap simulations of example 1 from hierarchy 1 assuming all stages are SRSWR

A simple example of how bootstrapping is performed: Assume vessels are randomly selected with replacement (WR) from a reference fleet. Trips from these vessels are then randomly sampled within each quarter (stratum). Fishing operations within each selected trip are randomly selected with replacement from multiple fishing operations that occurred on the trip. Within a fishing operation, fish are randomly selected for length measurements and assigned to length classes once measured. For each observed length class, a subset of fish is randomly selected for ageing. Note that this code is easily modified to allow for the case where fish are selected for ageing from a self-sample that is separate from the sample of fish measured for length.

In reality, no stage is actually sampled WR, but this bootstrapping approach could be used for SRSWOR at every stage if one desires a conservative estimate of the variance of the estimates, i.e. a variance estimate larger than the WOR sampling variance.

4.4.1 Basic bootstrapping steps for a single simulated sampling effort

21. Take an SRSWR of n vessels from the N vessels in the reference fleet.

Within each quarter:

22. For each vessel selected in step (1), take an SRSWR of $t_{v(q)}$ trips from the full set of $T_{v(q)}$ trips by the vessel in that quarter.
23. For each trip selected in step (2), take an SRSWR of $h_{t(v,q)}$ fishing operations from the full set of $H_{t(v,q)}$ fishing operations on that trip by that vessel:
    a) If estimates of discard weights for the quarter are desired, then perform required calculations.

If biological parameter estimates are desired:

24. (Figure 2) For each fishing operation selected in step (3), take a
    a) SRSWR of $fL_{h(t,v,q)}$ individual fish for length measurements in the landings and
    b) SRSWR of $fD_{h(t,v,q)}$ individual fish for length measurements in the discards
25. If Numbers at Length (NAL) are required, perform required fishing operation-level (or trip-level) calculations (depending on aggregation level).
    a) Store the NAL_q^b calculated in step (5) in a temporary file. The notation indicates that it is the estimated NAL for quarter q obtained from bootstrap simulation b

26. For each length class observed in step (4), take a
    a) SRSWR of $faL_{l,h(t,v,q)}$ individual fish for age/weight measurements in the landings and
    b) SRSWR of $faD_{l,h(t,v,q)}$ individual fish for age/weight measurements in the discards
27. For stratum-level (quarterly) Age-Length Key (ALK_q^b), use all data from step (5) combined over all trips within the stratum
28. For calculating the quarterly Numbers at Age (NAA) or mean length at age ($\overline{LAA}$), use the NAL estimate from step (5) raised to the appropriate level and the ALK estimate from step (7)
    a) Store the NAA_q^b calculated in step (8) in a temporary file. The notation indicates that it is the estimated NAA for quarter q obtained from bootstrap simulation b
    b) Store the mean length at age $\overline{LAA}_q^b$ from step (8) in a temporary file.
29. Repeat steps (2) to (8) for each quarter to obtain the estimated NAA_q^b, NAL_q^b, and $\overline{LAA}_q^b$ for every quarter.
    a) Calculate and store the bootstrap estimates of the annual NAA^b, NAL^b, and $\overline{LAA}^b$ using the usual stratified weighted mean of the quarterly estimates
30. Repeat steps (1) through (9), B times to obtain B estimates of NAA^b, NAL^b, and $\overline{LAA}^b$ and if desired the quarterly estimates NAA_q^b, NAL_q^b, and $\overline{LAA}_q^b$.
31. The B estimates from step (10) can be used to
    a) Assess bias within each quarter by calculating the mean of the bootstrapped NAA_q^b, $\overline{LAA}_q^b$, and NAL_q^b values from the B simulated samplings and comparing the means to the NAA_q, $\overline{LAA}_q$, and NAL_q calculated using the original dataset
    b) Assess bias of the annual estimates by calculating the mean of the bootstrapped NAA^b, NAL^b, and $\overline{LAA}^b$ values from the B simulated samplings and comparing the means to the NAA, $\overline{LAA}$, and NAL calculated using the original dataset
    c) Assess the standard errors (or variances or relative standard errors) of the original quarterly estimates of NAA_q, $\overline{LAA}_q$, and NAL_q by calculating the standard errors (or variance or relative standard errors) of the B estimates of NAA_q^b, NAL_q^b, and $\overline{LAA}_q^b$.
    d) Assess the standard errors (or variances or relative standard errors) of the original annual estimates of NAA, $\overline{LAA}$, and NAL by calculating the standard errors (or variance or relative standard errors) of the B estimates of NAA^b, NAL^b, and $\overline{LAA}^b$.
    e) Assess the correlation among the NAA or NAL by calculating the variance-covariance matrix for the desired vector. This is of use for determining the effect of the sampling strategy on the distribution of the vector of numbers at age or length (Note that a typical approach by a stock assessor is to consider the values in a NAA or NAL vector to be independent and distributed as a multinomial. The reality is they are not!).

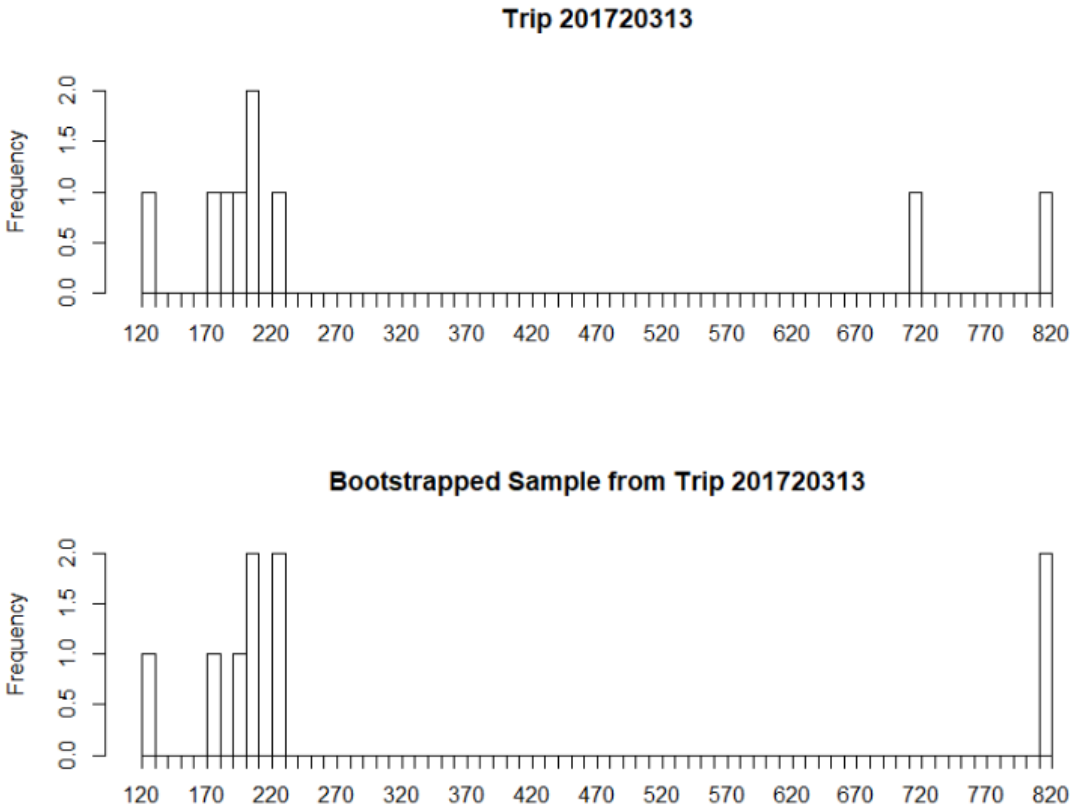


Figure 4.1.Example of sample data (top) of fish lengths measured during a single trip and one bootstrap SRSWR from that data (bottom).

4.5 Example code for bootstrapping a stratified multi-stage sampling design assuming all stages are SRSWR

In this example, there is a single fish species for which biological parameters, numbers at length and numbers at age, are desired by stratum. Strata are combinations of quarter, area and gear (indicated by "qX geartypeY areaZ") and all together there are five strata. There are three stages of sampling:

- vessel × trip (PSU) within strata;
- individual fish for length measurements on a trip (SSU) where only a single fishing operation was sampled;
- and length-stratified sub-sampling of fish (TSU) for aging and weight measurements.

The catch category (DIS or LAN) is also available and could have been included as a stratum at the SSU level but was not used in this example.

The bootstrapping R code in Box 2 in Annex A3.3 below assumes that there are two data files (Box 1) for this example: the first file ("fish") contains a record for each length class observed on each sampled trip of the number of fish observed in that length class on that trip and the second file ("indfish") contains individual fish records for age, length class, and weight for each trip where fish were sampled. The sampling for age was stratified by length class and the age subsample size was the maximum of 10 fish or the number of fish observed in that length class at

the second stage. This section is easily modified to allow for a separate self-sample of fish from a trip to be used for developing the ALK.

The example R code was developed to assess the bias and precision of the quarterly estimated NAL and NAA from the original datasets. In the code, these estimates are not expanded to the entire population or even to the sampling frame since data on total discard or landing weight by stratum was not available; instead, the results from this analysis provide the estimated NAL and NAA for the sampled trips only. Hence, many steps that are needed to expand the bootstrap estimates to the overall totals for the entire fishery are not shown. Note though that the expansion factors are constants and so it should not be difficult to include the additional computational steps. One need only have the values stored and accessible to the R program environment.

Box 1. A) Example of the trip level dataset "fish". B) Example of the subsample data for individual fish ages and weights "indfish".

A) Example of "fish" dataset

	NewStratum	CatchCat	UniqTrip	Length	SumStationWt	SumSampleWt	SumNumAtLen	RatioStaWtSampWt	TrpNumAtLen
1	q12Active22-24	DIS	201720302	220	0.312	0.097	1	3.216495	4
2	q12Active22-24	DIS	201720304	120	19.570	4.541	1	4.309623	5
3	q12Active22-24	DIS	201720304	140	19.570	4.541	2	4.309623	9
4	q12Active22-24	DIS	201720304	160	19.570	4.541	3	4.309623	13
5	q12Active22-24	DIS	201720304	170	19.570	4.541	3	4.309623	13
6	q12Active22-24	DIS	201720304	180	19.570	4.541	5	4.309623	22

B) Example of "indfish" dataset

	UniqTrip	CatchCat	Length	Age	Weight	NewStratum
1	201820301	DIS	210	2	83	q12Active22-24
2	201820301	LAN	360	3	461	q12Active22-24
3	201820301	LAN	370	2	510	q12Active22-24
4	201820301	LAN	370	2	471	q12Active22-24
5	201820301	LAN	380	2	565	q12Active22-24
6	201820301	LAN	380	2	579	q12Active22-24

4.6 Example code for bootstrap sampling of single-stage by SRSWOR

Suppose instead that at one or more stages in the last example, an SRSWOR is required. There are several different approaches to bootstrapping without replacement sampling designs (cf. Mashreghi *et al.*, 2016). One possible approach is to construct a pseudo-population at the appropriate level based on the method described by Sitter (1992). Some example R code is provided Box 3 in Annex A3.3 below.

In the example in Box 3, a population of N=350 values is created and a sample of n = 100 is taken by SRSWOR. The code includes how to create the pseudo-population and do bootstrapping to obtain means and variances. It also shows that using the Horvitz-Thompson estimator of the total inside of the bootstrapping loop leads to a biased estimate whereas calculating an estimate for the population total after bootstrapping is unbiased. This set of code also shows that one could simply take an SRSWR and adjust the variance estimate with the finite population correction factor after bootstrapping to obtain the same results as those based on the pseudo-population.

4.7 Example code for bootstrapping two-stage sampling by SRSWOR at each stage

In the example in Box 4 in Annex A3.4, a population of $N = 350$ PSUs is created and a sample of $n = 30$ is taken by SRSWOR. For each PSU, there are between 10 and 20 SSUs of which a sample of 5 is taken for every PSU. For our example, suppose a PSU is a vessel × trip and each PSU has between 10 and 20 hauls. On each trip, the observer samples 5 hauls for the weight of discards ($dw_{h(t,v)}$) and the weight of landings ($lw_{h(t,v)}$) of a single species of fish but has the total weight of the species for every haul, sampled or not ($W_{h(t,v)}$). Of interest is estimating the total discard weight of the species for the entire population of PSUs. The code shows how to create the pseudo-population and do bootstrapping to obtain the estimated total discard weight and its variance. For this example, we used the following method for estimating discard weight:

32. For each sampled haul, estimate the total discard weight:

$$\widehat{dW}_{h(t,v)} = \frac{dw_{h(t,v)}}{lw_{h(t,v)}} W_{h(t,v)}$$

The assumption here is that a haul is sampled by the observer before the discards and landings are separated during the processing of the haul by the crew. Hence, the proportion of the observer's sample that is discards is considered equivalent to the proportion of the total landed weight of the haul that is discards.

33. For each sampled vessel × trip the haul data are aggregated as follows. Total landings and discards for sampled hauls on vessel × trip $t(v)$ are defined as the sums of all sampled hauls

$$LW_{t(v)} = \sum_{h(t,v)} LW_{h(t,v)}$$

and

$$DW_{t(v)} = \sum_{h(t,v)} \widehat{dW}_{h(t,v)}.$$

The unsampled hauls are excluded from the raising procedure. Note that the total landings weight for a haul, $LW_{h(t,v)}$, is assumed to be known (e.g. provided by captain) or, if not, it can be calculated using $\widehat{lW}_{h(t,v)} = W_{h(t,v)} - \widehat{dW}_{h(t,v)}$ in place of $LW_{h(t,v)}$ in the above equation.

34. To obtain estimated total discards for the entire population of PSUs, vessel × trip-level estimates from step (2) are aggregated as follows. The estimated total discard weight, $\widehat{DW}$, is given by

$$\widehat{DW} = \frac{\sum_{t(v)\in S} DW_{t(v)}}{\sum_{t(v)\in S} LW_{t(v)}} \times LW$$

where the sums in the ratio are over all observed trips in S and $LW = \sum_{t(v)\in P} LW_{t(v)}$ is the total landed weight reported for all vessel × trips in P, observed or not.

4.8 Bibliography

Bickel, P. J. and D. A. Freedman (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics* 12(2), 470–482.

Chao, M. T. and S.-H. Lo (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. *Statistica Sinica* 4(2), 389–406

Efron, B. and R. Tibshirani, 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1(1), 54-75.

Mashreghi, Z., D. Haziza and C. Léger, 2016. A survey of bootstrap methods in finite population sampling. *Statistics Surveys,* 10, 1–52.

Sitter, R. R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics* 20 (2), 135–154.

5 Bias

5.1 Method

The main ICES reports that considered the topic of bias in relation to catch sampling programs are:

- WKACCU. ICES. 2008. Report of the Workshop on Methods to Evaluate and Estimate the Accuracy of Fisheries Data used for Assessment (WKACCU), 27–30 October 2008, Bergen, Norway. ICES CM 2008\ACOM:32. 41 pp.
- SGPIDS. ICES. 2011. Report of the Study Group on Practical Implementation of Discard Sampling Plans (SGPIDS), 27 June–1 July 2011, ICES Headquarters, Denmark. ICES CM2011/ACOM: 50. 116 pp.
- WKPICS. ICES. 2012. Report of the Working Group on Practical Implementation of Statistical Sound Catch Sampling Programs, 8–10 November 2011, Bilbao, Spain. ICES CM 2011/ ACOM:52. 55 pp.

The potential sources of bias that these reports identify were collated and an analysis was performed to see whether the RDBES could provide information that helped to evaluate that source of bias. The full analysis is presented in Annex A3.4 below. For specific cases where the RDBES can already provide insights into bias, some example reports have been developed and these are provided in Annex A3.5 below. In some cases where the RDBES is currently not capable of informing about the bias, some changes are suggested to the WGRDBESGOV[13] core group that could enhance the capacity of the system in bias analyses.

5.2 Summary

In the table below, there were 41 issues identified from the ICES literature referenced above, classified into six categories. It was found that the RDBES can already provide comprehensive information about 12 of these issues, and partial information about a further 18 issues. This is a great improvement over systems such as the RDB or InterCatch, which can provide little or no information about these potential sources of bias.

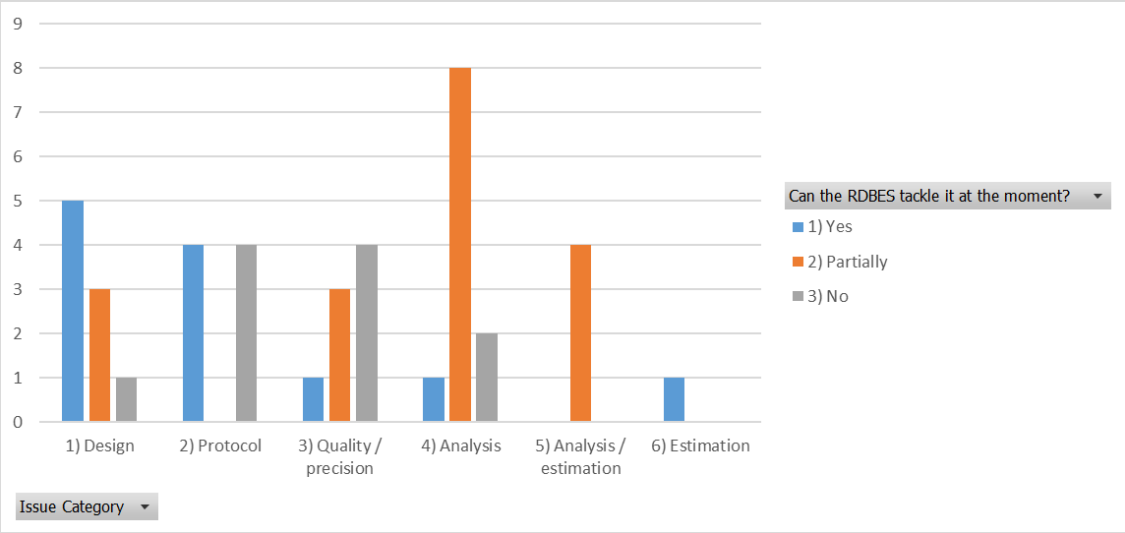
The RDBES could not provide information about 11 issues. Typically, this required information that would not normally be stored in a commercial fisheries database, such as detailed protocols or training records.

Table 5.1. Table 1 Can the RDBES inform about the bias issues identified?

	1) Yes	2) Partially	3) No	Total
1) Design	5	3	1	9
2) Protocol	4		4	8
3) Quality / precision	1	3	4	8
4) Analysis	1	8	2	11

[13] https://www.ices.dk/community/groups/Pages/WGRDBESGOV.aspx

	1) Yes	2) Partially	3) No	Total
5) Analysis / estimation	0	4	0	4
6) Estimation	1	0	0	1
Total	12	18	11	41



6 Future work

6.1 Introduction

The RDBES aims to tackle long-standing needs of commercial catch sampling and estimation within ICES in terms of both precision and bias. It is a collaborative effort that involves experts and many countries and that will take several years to develop (the RDBES development roadmap is reviewed and updated annually[14]). The tools to evaluate data precision and bias described within this report are necessarily limited in scope to ensure the work was feasible to complete with the time and resources available. They should not be considered as a complete solution but as a first step towards creating tools that can be applied more widely.

This section describes some ways in which this work could be extended. A roadmap is presented which shows how this could be approached. It will of course be necessary to gain feedback from users about the tools specified in this report and implement any identified improvements.

6.2 Analytical calculation of variance

The current work on implementing the analytical functions in R code was only done for totals so could be extended to include other estimates, such as numbers at length. It could also be extended to additional estimators (such as ratio estimators and age-length keys) within simple sampling designs (simple random sampling with or without replacement at each stage). Since the analytical calculation of variance is only feasible in relatively simple sampling programmes it is unlikely to be of much use when confronted with real-world data. For this reason it is not recommended to continue the development of these functions beyond the work done in this report.

6.3 Bootstrapping

6.3.1 Implementation in R code

Implement the bootstrapping pseudo-code as an R function which uses the correct RDBES field names. As far as possible the code should be modularized and split into functions to allow it to be more flexible.

6.3.2 Post-stratification and domain estimation

For this work, the bootstrapping approach is based on the estimation of parameters for the strata defined in the sampling design and reported in the data. However, often we want to produce estimates for domains that do not match the sampling strata. The bootstrapping procedure and the input datasets can be amended do this correctly. If the number of samples in strata are proportional to the size of the population (e.g. number of fishing trips) then the calculation of post-stratification variance based on the sampling strata will be reasonable, but in most cases, this will

[14] ICES. 2020. Steering Committee of the Regional Fisheries Database (SCRDB; outputs from 2019 meeting). ICES Scientific Reports. 2:24. 57 pp. http://doi.org/10.17895/ices.pub.5992.

not be the case and the post-stratification variance will actually be higher than that calculated based on the sampling strata.

6.3.3 Age-Length Key (ALK) construction

In the examples provided, the ALK is compiled across all samples to give an unweighted stratum-level ALK. Since the bootstrapping approach must honour the sampling design, the estimate of the accuracies associated with the numbers at age using this estimator are correctly captured. But the question of whether the current estimation procedure of simply collating all data collected within a stratum to construct the ALK provides the best estimator (more accurate as defined here) is unknown. The bootstrapping method could be used to investigate whether this is the most appropriate way to estimate the ALK or should a weighted approach be used instead[15].

6.3.4 Estimation based on other sampling designs

The procedures provided in this report are based on sampling designs that assume that the units in the sampling frame are selected based on simple random sampling, that is, every unit in the frame is equally likely to be the one selected. There are sampling efforts in MS that rely on unequal probability selection, usually based on a probability proportional to size, for example, assigning the likelihood of selecting a vessel to have an observer on board based on the vessel's total landings in the prior year. Hence, more active or larger vessels are more likely to be sampled than the "smaller" vessels. The bootstrapping approach described here can be modified to allow for unequal probability sampling.

6.3.5 Extension to other types of estimation

This work only considered national design-based estimation but there are other types to be considered:

- National design-based estimation using the RDBES clustering variables
- Regional design-based estimation without/with the RDBES clustering variables
- National model-based estimation without/with the RDBES clustering variables
- Regional model-based estimation without/with the RDBES clustering variables

Of these, the regional design-based estimation should be the priority since it is likely not difficult to extend the work to accommodate a regional approach, and this extension is relevant to the move towards designing regional sampling programmes.

6.4 Bias

There are a number of actions described in the table listing the potential sources of bias in Annex A3.4 below. When these actions are considered, priority should be given to those actions that have the most benefit but that do not require a change in the RDBES data model. Some of the most important issues raised are inherently informed by the existing RDBES data model and will not require any further development work. For example, the way that issue numbers 1 (sampling

[15] Sondre Aanes and Jon Helge Vølstad. Efficient statistical estimators and sampling strategies for estimating the age composition of fish. Canadian Journal of Fisheries and Aquatic Sciences. 72(6): 938-953. https://doi.org/10.1139/cjfas-2014-0408

design), 4 (spatial and temporal coverage), 5 (sample allocation schemes), and 8 (PSU selection) are considered in the future will be improved by sampling data being uploaded to the RDBES.

6.5 Roadmap

A suggested roadmap for further development of the topics raised above is presented below. This is presented as a guide, assuming both funding and expert time will be available. Still, it will be necessary to further quantify the resources required and identify if those resources can be made available. Since the development of estimation methods for the RDBES is ongoing any further work should be considered within the overall RDBES roadmap and the work plan of related groups such as the EU Regional Coordination Groups (RCGs).

	2021 Q1	2021 Q2	2021 Q3	2021 Q4	2022 Q1	2022 Q2
Implement bootstrapping as R function						
Extend bootstrapping to regional sampling						
Bootstrapping for post-stratification and domain estimation						
Development of further example reports for potential bias						
WGRDBESGOV Core Group to consider the identified actions that are related to bias						

7 Summary

This report shows how the new RDBES that is currently in development will be better able to support the recast EU Data Collection Framework (Regulation (EU) 2017/1004) than the existing RDB. The RDBES is an essential platform for MS and RCGs to fulfil their obligations towards documenting and improving data quality and designing and implementing regional sampling designs.

The evaluation of data precision was performed using two complementary techniques. For relatively simple sampling designs it is possible to use analytical functions to calculate the precision (or a related statistical measure such as variance) of a statistical estimate. These calculations and implementations of them in R code are presented in this report. For more complicated sampling designs, the use of analytical functions is usually not feasible. In these cases, it is necessary to evaluate precision using numerical techniques, the main one of which is bootstrapping. This report discussed when bootstrapping is appropriate and gives several worked examples describing how bootstrapping can be applied in different cases.

The evaluation of bias is a difficult subject and is hard to quantify. The approach followed in this report was to build on the previous work available in the ICES literature and identify and enumerate the main common sources of bias in catch sampling programs they describe. The information was collated and an evaluation performed as to whether data stored using the RDBES data format and reports issues from them can inform about the potential for bias in catch estimates. A set of example reports was coded that demonstrates the utility of the RDBES in relation to bias issues and can already help member states to identify how deviations in their sampling programmes and sampling variability may potentially lead to bias in their catch estimates.

Annex A3.1 List of participants

Name	Institute	Country	Email
Mary Christman		USA	marycchristman@gmail.com
Liz Clarke	Marine Scotland Science	United Kingdom	liz.clarke@gov.scot
David Currie	Marine Institute	Ireland	david.currie@marine.ie
Annica Isaksson de Groote	SLU	Sweden	annica.isaksson.de.groote@slu.se
Kirsten Birch Håkansson	DTU Aqua	Denmark	kih@aqua.dtu.dk
Nuno Prista	SLU	Sweden	nuno.prista@slu.se
Jose Rodriguez	IEO	Spain	jose.rodriguez@ieo.es

Annex A3.2 R implementation of analytical variance calculation using design-based estimation

The following code creates a function that acts on an R object containing sampling data, and calculates a Horvitz-Thompson estimate of the population total of a univariate variable (for example landing weight) and the associated variance for that estimate, assuming simple random sampling without replacement (SRSWOR) was used at each stage. The code has been tested for a 3-stage sampling design and can be found on Github[16].

```
estimateHTtotalMultiStageSRSWOR <- function(RDBobj, stages=stages,
varOfInterest="SAtotalWtLive"){
  # this function calculates, and outputs, the Horvitz-Thompson estimate
  # of the population total of a single univariate variable
  # and the resulting variance of the estimate, assuming
  # SRSWOR (simple random sampling without replacement) is used at each stage
  # The function has 3 arguments:
  # RDBobj - the object containing the data, in RDBES format
  # stages - the sampling stages in the data
  # this avoids the need to specify hierarchies used in RDBES
  # varOfInterest - a character string specifying the name of the variable
  # for which we are estimating the population total (the "y variable")

  # set up some objects
  nStage <- length(stages)
  idPrev <- idList <- piList <- list()
  y <- nTotal <- nSamp <- meanStage <- ymean <- list()
  ssqTerm <- ssqStage <- tStage <- nTotStage <- nSampStage <- list()
  y[[nStage]] <- RDBobj[[stages[[nStage]]]][,varOfInterest]
  estVarTot <- vv <- list()

  # create lists of the key variables required in the calculations
  # nTotal - total number of units in each stage
  # nSamp - total number of samples in each stage
  # idList - the unique identifier for each unit in each stage
  # idPrev - a unique identifier for the units in the previous stage
  # at each stage (except stage 1)
  for (i in 1:nStage) {
    dat <- RDBobj[[stages[[i]]]]
    nTotal[[i]] <- dat[,paste(stages[[i]],"numTotal",sep="")]
    nSamp[[i]] <- dat[,paste(stages[[i]],"numSamp",sep="")]
    idList[[i]] <- dat[,paste(stages[[i]],"id",sep="")]
    if (i>1) {
      idPrev[[i]] <- dat[,paste(stages[[i-1]],"id",sep="")]
    } # end if
  } # end for

  # calculate terms in the variance at each stage
```

[16] https://github.com/ices-eg/WK_RDBES/tree/master/Special_Request_20_05

```
  # this needs to be done from the lowest heirarchy upwards
  # meanStage is the mean of of the y variable at each stage
  # nTotStage & nSampStage are the numbers of units in each stage
  # ssqStage is the sum of squares term in the variance estimate
  # tStage the estimate of the "population" total of the y variable
  # at each stage
  for (i in nStage:1) {
    if (i==1) {
      # for the first stage (which is calculated last) we use sum not tapply
      meanStage[[i]] <- mean(y[[i]])
      ymean[[i]] <- rep(meanStage[[i]],length(idList[[i]]))
      nTotStage[[i]] <- rep(mean(nTotal[[i]]),length(idList[[i]]))
      nSampStage[[i]] <- rep(mean(nSamp[[i]]),length(idList[[i]]))
      ssqStage[[i]] <- sum(((y[[i]]-ymean[[i]])^2)/(nSamp[[i]]-1))
      tStage[[i]] <- sum(nTotal[[i]]/nSamp[[i]]*y[[i]])
    } else {
      # as there are several units of upper hierarchies in each stage,
      # we use tapply
      meanStage[[i]] <- tapply(y[[i]],idPrev[[i]],mean)
      ymean[[i]] <- meanStage[[i]][match(idPrev[[i]],names(meanStage[[i]]))]
      nTotStage[[i]] <- tapply(nTotal[[i]],idPrev[[i]],mean)
      nSampStage[[i]] <- tapply(nSamp[[i]],idPrev[[i]],mean)
      ssqStage[[i]] <- tapply(((y[[i]]-ymean[[i]])^2)/
           (nSamp[[i]]-1),idPrev[[i]],sum)
      tStage[[i]] <- tapply(nTotal[[i]]/nSamp[[i]]*y[[i]],idPrev[[i]],sum)
    }
    # add in cases where variance is zero because the whole population
 at that stage was sampled
    ssqStage[[i]][is.infinite(ssqStage[[i]]) &
      nTotStage[[i]]==nSampStage[[i]]] <- 0
    ssqTerm[[i]] <- nTotStage[[i]]^2*(1-nSampStage[[i]]/
      nTotStage[[i]])/nSampStage[[i]]*ssqStage[[i]]
    if (i>1) y[[i-1]] <- tStage[[i]][match(idList[[i-1]],names(tStage[[i]]))]
  } #end for
  estTot <- tStage

  # now calculate the sum of each term sequentially from the lowest hierarchy
  # to the first heirarchy
  vv <- estVarTot <- list()
  vv[[nStage]] <- 0
  for (i in nStage:2) {
    idPrevVec <- idList[[i-1]][match(names(ssqTerm[[i]]),idList[[i-1]])]
    estVarTot[[i]] <- tapply(ssqTerm[[i]]+vv[[i]],idPrevVec,sum)
    vv[[i-1]] <- nTotStage[[i]]/nSampStage[[i]]*estVarTot[[i]]
  }

  # output the point estimate of the population total, and
  # the associated variance estimate for the point estimate
  output <- list(estTot=estTot,estVarTot=estVarTot)
  return(output)
}
x
```

Annex A3.3 Bootstrapping pseudo-code

Box 2. R code for bootstrapping two stage sampling of trips and then fish within trips. It includes stratified SRSWR of fish for age where the strata are length classes observed in a trip.x

Box 2. R code for SRSWR of several stages

```
# PREPARATORY STEPS
# Number of bootstraps desired for simulating results
TB <- 2500

# List of strata names
stratalist <- unique(fish$NewStratum)

# matrices to store stratum-level NAL results (means and variances) from the bootstraps
stratbootNAL <- matrix(0, nrow=length(unique(fish$Length)), ncol=length(stratalist))
rownames(stratbootNAL) <- sort(unique(fish$Length))
colnames(stratbootNAL) <- stratalist
stratbootrtvarNAL <- matrix(0, nrow=length(unique(fish$Length)), ncol=length(stratalist))
rownames(stratbootrtvarNAL) <- sort(unique(fish$Length))
colnames(stratbootrtvarNAL) <- stratalist

# matrices to store stratum-level NAA results (means and variances) from the bootstraps
stratbootNAA <- matrix(0, nrow=length(unique(indfish$Age)), ncol=length(stratalist))
rownames(stratbootNAA) <- sort(unique(indfish$Age))
colnames(stratbootNAA) <- stratalist
stratbootrtvarNAA <- matrix(0, nrow=length(unique(indfish$Age)), ncol=length(stratalist))
rownames(stratbootrtvarNAA) <- sort(unique(indfish$Age))
colnames(stratbootrtvarNAA) <- stratalist

# need column indicator for stratum ID
colid <- 0

#### OUTSIDE LOOP FOR STRATUM ANALYSES
for (strata in stratalist)
{
  # matrix to store NALs from each bootstrap sample within a stratum
```

Box 2. R code for SRSWR of several stages

```
bootNAL <- matrix(0, nrow=length(unique(fish$Length)), ncol=TB)
rownames(bootNAL) <- sort(unique(fish$Length))
bootNAA <- matrix(0, nrow=length(unique(indfish$Age)), ncol=TB)
rownames(bootNAA) <- sort(unique(indfish$Age))

# create a temporary dataset of lengths for the stratum (bootstrap samples taken from this)
tempdata <- fish[fish$NewStratum == strata, ]

# need a list of trips in tempdata
trips <- unique(tempdata$UniqTrip)
# need total number of trips in tempdata
lentrips <- length(trips)

##### BOOTSTRAPPING LOOP (within a stratum)
for (boots in 1:TB)
{
# for each bootstrap need 2 datasets to store the data from the bootstrapped trips
# one for the length sampling and one for the subsampling of ages by length

# matrix to store bootstrap samples of trips for length measurements
# first row will be removed later – it contains all 0s
bootfish <- matrix(0, nrow=1, ncol=dim(fish)[2])
colnames(bootfish) <- colnames(fish)

# matrix to store bootstrap subsamples of age data from bootstrapped trips
# first row will be removed later – it contains all 0s
bootind <- matrix(0, nrow=1, ncol=dim(tempind1)[2])
colnames(bootind) <- c("UniqTrip", "CatchCat", "Length", "Age", "Weight", "NewStratum")

# take a SRSWR of trips within the stratum using SRSWR
boottrip <- sample(trips, lentrips, replace = T)

####### TRIPS LOOP
for (tripid in boottrip)
```

Box 2. R code for SRSWR of several stages

```
{
  # fish lengths from a sampled tripid
  bootdata <- tempdata[tempdata$UniqTrip==tripid,]
  rownames(bootdata) <- bootdata$Length

  # total number of fish originally sampled on the trip
  numlens <- sum(bootdata$SumNumAtLen)

  # relative frequencies of each length in dataset
  Sprobs <- bootdata$SumNumAtLen/sum(bootdata$SumNumAtLen)

  # TAKE A BOOTSTRAP SRSWR OF FISH WITHIN THE TRIP FOR LENGTH MEASUREMENTS
  fishsamp <- sample(bootdata$Length, numlens, replace=T, prob=Sprobs)

  # frequencies of fish lengths in the bootstrap sample of fish
  numtimes <- table(fishsamp)

  # create new fish dataset (contains only 1 record for each length) for the trip
  tempfish <- bootdata[which(!is.na(match(rownames(bootdata),fishsamp))),]
  # order it by lengths
  tempfish <- tempfish[order(tempfish$Length,decreasing=FALSE),]

  # update the frequencies of fish sampled in bootstrapped trip by length
  tempfish$SumNumAtLen <- numtimes

  # calculate new estimated numbers per trip (using the ratio of sample to total weight
 # within the  catch category
  tempfish$TrpNumAtLen<-ceiling(tempfish$SumNumAtLen*tempfish$RatioStaWtSampWt)

  # save bootstrapped trip of fish length data to temporary file (used for bootstrapped NAL)
  bootfish <- rbind(bootfish, tempfish)

  # obtain the data from indfish for the same tripid
```

Box 2. R code for SRSWR of several stages

```
  tempind <- indfish[indfish$UniqTrip==tripid,]

  # list of lengths in the bootstrap sample of fish in trip = tripid
  indlens <- unique(tempfish$Length)

  # SRSWR STRATIFIED BY LENGTH CLASS OF FISH AGES USING LENGTHS FROM TEMPFISH
  for (j in indlens)
    {
      # number of times a length appears in fish dataset
      size1 <- tempfish$SumNumAtLen [tempfish$Length == j])
      # take a SRSWR of size1 or 10, which ever is smaller from indfish
      samp1 <- sample(min(10, size1), replace=T)
      subsample1 <- tempind[tempind$Length == j,]
      # store results for that length class
      bootind <- rbind(bootind, subsample1[samp1,])
      } # close j loop
} # close tripid loop

# ANALYZE the "boots" BOOTSTRAP SAMPLE AND STORE RESULTS
# the two datasets bootfish and booind contain "new" trips that have the same numbers of
#  fish for length measurements but may contain different numbers of fish for the age
#  measurements

# remove the first row of the bootfish and bootind datasets (all zeroes)
totfishrecs <- dim(bootfish)[1]
if (totfishrecs> 1)  bootfish <- bootfish[2:totfishrecs,]
totindrecs <- dim(bootind)[1]
if (totindrecs > 1) bootind <- bootind[2:totindrecs,]

# store NAL for the "boots" bootstrap
numlen <- tapply(bootfish$TrpNumAtLen, bootfish$Length, sum)
bootNAL[which(!is.na(match(rownames(bootNAL),names(numlen)))), boots] <- numlen

# calculate NAA for the "boots" bootstrap
```

Box 2. R code for SRSWR of several stages

```
 # construct ALK for the "boots" bootstrap
 bootALK <- table(bootind$Length, bootind$Age, useNA="no")
 bootALK <- bootALK/apply(bootALK, 1, sum)
 # multiply ALK by NAL for the "boots" bootstrap
  numagelen <- sweep(as.matrix(bootALK[which(!is.na(match(rownames(as.matrix(boot-
ALK)),rownames(as.matrix(bootNAL)))))),]), MARGIN=1, as.matrix(boot-
NAL[which(!is.na(match(rownames(as.matrix(bootNAL[,1])),rownames(as.matrix(boot-
ALK))))),boots]), '*')
  # collapse to get NAA and store
  bootNAA[which(!is.na(match(rownames(bootNAA),rownames(t(numagelen))))),boots] <-
apply(numagelen, 2, sum)
} # close boots loop

# store the means and variances of the bootstrapped NAA and NAL for stratum = "strata"
colid <- colid + 1
stratbootrtvarNAL[,colid]<- sqrt(apply(bootNAL, 1, var))
stratbootNAL[,colid]    <- apply(bootNAL, 1, mean)
templist <- apply(bootNAA,2, sum)
tempbootNAA <- bootNAA[,templist>0]
stratbootrtvarNAA[,colid]<- sqrt(apply(tempbootNAA, 1, var))
stratbootNAA[,colid]    <- apply(tempbootNAA, 1, mean)
} # close strata loop
```

Box 3. Simple example of constructing and bootstrapping from a pseudo-population using the approach of Sitter (1992).

Box 3. R code for Sitter (1992)

```
################################## Sitter 1992 approach ##################################

##################### Pseudo-Population when N/n is not an integer #####################

##################### Random size of Pseudo-Pop is not a problem for ###################

########## variance estimation, i.e. can be ignored, but could affect estimation of #############

######################### population total if pseudo-pop size is used ####################

# create a population of N = 350 values

N <- 350

pop <- rnorm(N,500,75)

POPmean <- mean(pop)

POPtotal <- sum(pop)


# take a SRSWOR from the population

n <- 100

SRSWOR <- sample(pop, n, replace=F)

# estimate of the population mean and its variance

SRSWORmean <- mean(SRSWOR)

SRSWORvar <- (1-(n/N))*var(SRSWOR)/n

# estimate of the population total and its variance

SRSWORht <- N*SRSWORmean

SRSWORhtvar = N^2 * SRSWORse^2


# calculations needed for creating the pseudo-population

f <- n/N

k <- (N/n)*(1-((1-f)/n))

k1 <- floor(k)

k2 <- ceiling(k)
```

Box 3. R code for Sitter (1992)

```
n1prime <- n-1
n2prime <- n
a1 <- (k1*(1-(n1prime/(n*k1))))/(n1prime*(n*k1-1))
a2 <- (k2*(1-(n2prime/(n*k2))))/(n2prime*(n*k2-1))
qs <- (((1-f)/(n*(n-1)))-a2)/(a1-a2)
PPB1 <- rep(SRSWOR, k1)
PPB2 <- rep(SRSWOR, k2)
len1 <- length(PPB1)
len2 <- length(PPB2)

B <- 50000
# place to store bootstrapping results
BootRes1 <- matrix(0, nrow=B, ncol=6)
colnames(BootRes1) <- c("SRSWORmean", "SRSWORvar", "HTBtotal", "HTBbias",
                                        "SRSWRmean", "SRSWRvar*(1-f)")

# bootstrapping loop
for (b in 1:B)
{
 #take a sample from the appropriate pseudo-population
 rannumber <- runif(1)
 if (rannumber <  qs)
 {
  sampWOR <- sample(PPB1, n1prime, replace=F)
  poplen <- len1
 }
 if (rannumber >= qs)
 {
  sampWOR <- sample(PPB2, n2prime, replace=F)
  poplen <- len2
 }
 BootRes1[b,1] <- mean(sampWOR)
```

Box 3. R code for Sitter (1992)

```
  BootRes1[b,2] <- (1-(n/N))*var(sampWOR)/n
  BootRes1[b,3] <- poplen*BootRes1[b,1]
  BootRes1[b,4] <- BootRes1[b,3] - BootRes1[b,1]*N
  # compare to SRSWR
  tempsamp <- sample(SRSWOR, n, replace=T)
  BootRes1[b,5] <- mean(tempsamp)
  BootRes1[b,6] <- var(tempsamp)*(1-f)/n
}



#Original Population
POPmean
POPtotal
#Original Sample
SRSWORmean
SRSWORvar
SRSWORht
# Means of bootstrap quantities
apply(BootRes1, 2, mean)
# variances of bootstrap quantities
apply(BootRes1, 2, var)

#SOME RESULTS
> # Population
> POPmean =       500.6
> POPtotal   = 175224.7

> # Original Sample
> SRSWORmean    =       499.18
> SRSWORvar       =         35.28
> SRSWORht         =  174712.50
```

Box 3. R code for Sitter (1992)

```
>
> # Means of bootstrap quantities (should reproduce the values calculated from the original sample)
> apply(BootRes1, 2, mean)
   SRSWORmean      SRSWORvar       HTBtotal        HTBbias      SRSWRmean SRSWRvar*(1-f)
      499.169         35.278     178925.23        4216.19        499.228         34.894
>
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
# Population of N = 350 PSUs
Npsu <- 350
# Each PSU has between 10 and 20 SSUs
SSUs <- sample(10:20, Npsu, replace=T)
# Total Number of SSUs for the Npsu PSUs
totnumSSU <- sum(SSUs)

# Create data for each SSU
# True Total Haul Weight
SSUtotwts <- sample(200:700, totnumSSU, replace=T)
# True proportion of total haul weight that is discards
SSUdisprops <- runif(totnumSSU, min=0.05, max=0.20)

# Create the population
pop <- data.frame(matrix(0, nrow=SSUs[1], ncol = 11))
colnames(pop) <- c("PSUid", "SSUid", "totSSU", "SSUtotwt", "SSUdisprop",
                   "trueSSUdiswt", "trueSSUlanwt", "SSUsampwt", "sampSSUdiswt",
"sampSSUlanwt",
                   "ratioSampDisLan")

# Fill in matrix for the first PSU
pop[,1] <- 1
pop[,2] <- c(1:SSUs[1])
pop[,3] <- SSUs[1]
pop[,4] <- SSUtotwts[1:SSUs[1]]
pop[,5] <- SSUdisprops[1:SSUs[1]]

# Do the same for the remaining PSUs
for (i in 2:Npsu)
{
  tpop <- data.frame(matrix(0, nrow=SSUs[i], ncol = 11))
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
 colnames(tpop) <- c("PSUid", "SSUid", "totSSU", "SSUtotwt", "SSUdisprop",
                     "trueSSUdiswt", "trueSSUlanwt", "SSUsampwt", "sampSSUdiswt",
"sampSSUlanwt",
                     "ratioSampDisLan")
 tpop[,1]<- i
 tpop[,2] <- c(1:SSUs[i])
 tpop[,3] <- SSUs[i]
 tpop[,4] <- SSUtotwts[1:SSUs[i]]
 tpop[,5] <- SSUdisprops[1:SSUs[i]]
 pop <- rbind(pop, tpop)
}
# Calculate weights for discards and landings for each SSU
pop[,6] <- pop[,4]*pop[,5]
pop[,7] <- pop[,4]-pop[,6]

# Create sample weights of discards and landings
pop[,8]  <- runif(totnumSSU, min=0.1, max=0.15)*pop[,4]
pop[,9]  <- pop[,8]*runif(totnumSSU, min=0.05, max=0.20)
pop[,10] <- pop[,8]-pop[,9]
pop[,11] <- pop[,9]/pop[,10]

# Parameter Values
TrueTotDisWt <- sum(pop[,6])
TrueTotLanWt <- sum(pop[,7])
TrueTotWt    <- sum(pop[,4])

# Population Estimate of Total Discards (can never be known)
# Might be useful for estimating bias due to using a ratio estimator
EstTripDis <- matrix(0, nrow=Npsu, ncol=2)
colnames(EstTripDis) <- c("EstDisWt", "SumLanWt")
for (i in 1:Npsu)
{
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
 PSUdata <- pop[pop$PSUid==i,]
 EstTripDis[i,1] <- sum(PSUdata$ratioSampDisLan*PSUdata$SSUtotwt)
 EstTripDis[i,2] <- sum(PSUdata$trueSSUlanwt)
}
EstPopTotDisWt <- TrueTotLanWt*sum(EstTripDis[,1])/sum(EstTripDis[,2])



# take a SRSWOR of PSUs from the population
npsu <- 30
SRSWORpsu <- sample(1:Npsu, npsu, replace=F)
PSUsample <- pop[which(!is.na(match(pop$PSUid,SRSWORpsu))),]

# for each sampled PSU, take a SRSWOR of nssu SSUs
nssu <- 5
temppop <- pop[pop$PSUid == SRSWORpsu[1],]
SRSWORssu <- sample(1:temppop$totSSU[1], nssu, replace=F)
PSUSSUsamp <- temppop[which(!is.na(match(temppop$SSUid,SRSWORssu))),]

for (i in SRSWORpsu[2:npsu])
{
 temppsu <- pop[pop$PSUid == i,]
 SRSWORssu <- sample(1:temppsu$totSSU[1], nssu, replace=F)
 tempssu <- temppsu[which(!is.na(match(temppsu$SSUid,SRSWORssu))),]
 PSUSSUsamp <- rbind(PSUSSUsamp, tempssu)
}

# Original Sample Estimate of Total Discards
EstTripDis <- matrix(0, nrow=Npsu, ncol=2)
colnames(EstTripDis) <- c("EstDisWt", "SumLanWt")
rowid <- 0
for (i in SRSWORpsu)
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
{
 PSUSSUdata <- PSUSSUsamp[PSUSSUsamp$PSUid==i,]
 rowid <- rowid + 1
 EstTripDis[rowid,1] <- sum(PSUSSUdata$ratioSampDisLan*PSUSSUdata$SSUtotwt)
 # Using true landings weight for the hauls not an estimated weight
 EstTripDis[rowid,2] <- sum(PSUSSUdata$trueSSUlanwt)
}
# Using the reported true total landings weight for the entire population
EstSampTotDisWt <- TrueTotLanWt*sum(EstTripDis[,1])/sum(EstTripDis[,2])

# Calculations needed for creating the pseudo-population of PSUs used in bootstrapping
fp <- npsu/Npsu
kp <- (Npsu/npsu)*(1-((1-fp)/npsu))
k1p <- floor(kp)
k2p <- ceiling(kp)
n1primep <- npsu-1
n2primep <- npsu
a1p <- (k1p*(1-(n1primep/(npsu*k1p))))/(n1primep*(npsu*k1p-1))
a2p <- (k2p*(1-(n2primep/(npsu*k2p))))/(n2primep*(npsu*k2p-1))
qsp <- (((1-fp)/(npsu*(npsu-1)))-a2p)/(a1p-a2p)
PPB1p <- rep(SRSWORpsu, k1p)
PPB2p <- rep(SRSWORpsu, k2p)
len1p <- npsu*k1p
len2p <- npsu*k2p


# Number of bootstrap samples to use
B <- 50000
# place to store bootstrapping results
BootRes1 <- matrix(0, nrow=B, ncol=1)
colnames(BootRes1) <- c("EstTotDis")
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
# bootstrapping loop
for (b in 1:B)
{
 # sample PSUs from the pseudo-population
 rannumber <- runif(1)
 if (rannumber <  qsp)
 {
  sampWORp <- sample(PPB1p, n1primep, replace=F)
  psulen <- len1p
 }
 if (rannumber >= qsp)
 {
  sampWORp <- sample(PPB2p, n2primep, replace=F)
  psulen <- len2p
 }
 # Calculations needed to estimate total discards in population
 EstTripDis <- matrix(0, nrow=length(sampWORp), ncol=2)
 colnames(EstTripDis) <- c("EstDisWt", "SumLanWt")

 # sample SSUs from sampled PSUs
 rowid <- 0
 for (i in sampWORp)
 {
  Nssu <- unique(pop$totSSU[pop$PSUid==i])
  fs <- nssu/Nssu
  ks <- (Nssu/nssu)*(1-((1-fs)/nssu))
  k1s <- floor(ks)
  k2s <- ceiling(ks)
  n1primes <- nssu-1
  n2primes <- nssu
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
a1s <- (k1s*(1-(n1primes/(nssu*k1s))))/(n1primes*(nssu*k1s-1))
a2s <- (k2s*(1-(n2primes/(nssu*k2s))))/(n2primes*(nssu*k2s-1))
qss <- (((1-fs)/(nssu*(nssu-1)))-a2s)/(a1s-a2s)
SRSWORssu <- unique(PSUSSUsamp$SSUid[PSUSSUsamp$PSUid==i])
PPB1s <- rep(SRSWORssu, k1s)
PPB2s <- rep(SRSWORssu, k2s)
len1s <- length(PPB1s)
len2s <- length(PPB2s)
rannumbers <- runif(1)
if (rannumbers <  qss)
{
 sampWORs <- sample(PPB1s, n1primes, replace=F)
 poplens <- len1s
}
if (rannumbers >= qss)
{
 sampWORs <- sample(PPB2s, n2primes, replace=F)
 poplen <- len2s
}
rowid <- rowid + 1
if (i == sampWORp[1])
{
 tempboot <- PSUSSUsamp[PSUSSUsamp$PSUid==i,]
 tempboot1 <- tempboot[match(sampWORs, tempboot$SSUid),]
}
if ( i !=sampWORp[1])
{
 tempboot <- PSUSSUsamp[PSUSSUsamp$PSUid==i,]
 tempboot1 <- tempboot[match(sampWORs, tempboot$SSUid),]
}
# EstTripDis now contains the data for the trips in this bootstrap sample
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
   EstTripDis[rowid,1] <- sum(tempboot1$ratioSampDisLan*tempboot1$SSUtotwt)
   EstTripDis[rowid,2] <- sum(tempboot1$trueSSUlanwt)
 }
# store estimate of total discards from the bootstrap sample
BootRes1[b,1] <- TrueTotLanWt*sum(EstTripDis[,1])/sum(EstTripDis[,2])
}

#Original Population Discard Weight
TrueTotDisWt
# Estimated Discard Weight Based on All Population Data
EstPopTotDisWt
# Estimated Discard Weight Based on Sample Data
EstSampTotDisWt
# Mean of the Bootstrap Estimates based on SRSWOR
mean(BootRes1[,1])
# Range of values of the Bootstrap Estimates
quantile(BootRes1[,1])
# Standard Error of the Bootstrap Estimates
sqrt(var(BootRes1[,1]))
# Relative Bias of mean of Bootstrap estimates compared to the sample used to create the pseudo-population
100*EstSampTotDisWt/mean(BootRes1)
# Confidence Interval Endpoints (90 and 95% CIs)
quantile(BootRes1[,1], probs=c(0.025, 0.05, 0.95, 0.975))

#####                    SOME                    RESULTS
#################################################################
> #Original Population Discard Weight
> TrueTotDisWt
[1] 312805.9
> # Estimated Discard Weight Based on All Population Data (had a sample been taken from every PSU)
```

Box 4. R code for two stage with SRSWOR at both stages. Estimating total discards in the population.

```
> EstPopTotDisWt
[1] 351515.4
> # Estimated Discard Weight Based on Sample Data of 30 PSUs and 5 SSUs/PSU
> EstSampTotDisWt
[1] 353891.9
> # Mean of the Bootstrap Estimates based on SRSWOR
> mean(BootRes1[,1])
[1] 355803.7
> # Range of values of the Bootstrap Estimates
> quantile(BootRes1[,1])
     0%      25%      50%      75%     100%
284957.9 344167.4 355730.1 367227.4 423707.4
> # Standard Error of the Bootstrap Estimates
> sqrt(var(BootRes1[,1]))
[1] 17013.32
# Relative Bias of mean of Bootstrap estimates compared to the sample used to create the pseudo-population
> 100*EstSampTotDisWt/mean(BootRes1)
[1] 99.46268
> # Confidence Interval Endpoints (90 and 95% CIs)
> quantile(BootRes1[,1], probs=c(0.025, 0.05, 0.95, 0.975))
   2.5%       5%      95%     97.5%
322741.0 328053.6 384104.4 389356.9
```

Annex A3.4 Potential sources of bias

The table below summarizes a number of potential sources of bias identified in the ICES literatures and describes if and how the RDBES can provide information about them. The columns have the following meanings:

- Issue Category
  - The broad category the issue falls within
- Id
  - A numerical id only used within this report
- Issue description
  - Text describing the potential source of bias and its effects. In the majority of cases descriptions were copied directly from the source.
- Source
  - The report where the issue was identified. The reports considered were:
    - WKACCU. ICES. 2008. Report of the Workshop on Methods to Evaluate and Estimate the Accuracy of Fisheries Data used for Assessment (WKACCU), 27–30 October 2008, Bergen, Norway. ICES CM 2008\ACOM:32. 41 pp.
    - SGPIDS. ICES. 2011. Report of the Study Group on Practical Implementation of Discard Sampling Plans (SGPIDS), 27 June - 1 July 2011, ICES Headquarters, Denmark. ICES CM2011/ACOM: 50. 116 pp
    - WKPICS. ICES. 2012. Report of the Working Group on Practical Implementation of Statistical Sound Catch Sampling Programs, 8 - 10 November 2011, Bilbao, Spain. ICES CM 2011/ ACOM:52. 55 pp.
- Affects
  - Which variables does this source of bias affect (e.g. age)? Some issues will affect all variables
- Can the RDBES tackle it at the moment?
  - Yes/No/Partially. Includes a description of how the RDBES can provide information about the issue
- Could the RDBES potentially tackle this in the future?
  - Could the RDBES potentially give information about this issue in the future? This might require a change in the RDBES data format.
- Comment
  - General comments related to the issue
- Actions
  - Actions required by specific groups to make progress on the issue

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
Design	1	Sampling design. Minimization of bias through sampling design, or at least an ability to identify and quantify biases, is more critical than minimization of variance (SGPIDS).	SGPIDS, WKPICS, 2011	All variables	**Yes** In the RDBES sampling design is documented alongside the data collected. It is then possible to evaluate whether the design might be a source of bias (e.g. non-probabilistic sampling, systematic non-responses)	-	Groups such as WGCATCH promote good practice in sampling design. Groups like WGRDBESGOV promote good practices in populating the RDBES format. Once fully populated the RDBES will be a valuable tool to analyse practical implementation of sampling designs.	-
	2	Coverage, design. If only part of the population is covered, the frame has under-coverage that will lead to bias unless the variables of interest (e.g. discard rates; species or size compositions) are the same in the parts of the population covered or not covered, or if only a very small part of the population is not covered.	WKPICS, 2011	All variables	**Partially** We can see what we have The RDBES stores population data and sampling data – these can then be compared to determine coverage. The RDBES does not store sampling frame data.	-	The overall sampling frames are described in the national work plans, and this is needed to evaluate the overall design. The RDBES will be able to support evaluation of national work-plans in the future.	-
	3	Coverage, country. An example of under-coverage would be the non-sampling of vessels of a national fleet that land in another country. This fraction may vary from year to year leading to a variable bias if activities, gears etc. differ from vessels landing in the home country.	WKPICS, 2011	All variables	**Partially** Using data in the RDBES it will be possible to quantify the magnitude of non-national vessel activity, but it is not possible to evaluate if it is included in a sampling frame.	-	National workplans provide detailed information about the sampling frames used. The recommended way of handling this situation within the RDBES needs reviewing.	Further analysis and discussion is possible at the Regional Coordination Groups and in the future when designing regional sampling plans. This issue will also be discussed and prioritized by the WGRDBESGOV "Core group".

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	4	Spatial and Temporal coverage: it has been discussed during the workshop that any discrepancy between the sampling and fishing effort coverage do not lead to a bias when the sampling is done randomly following a well-designed protocol. In other cases, the temporal coverage in terms of mean discrepancy between proportion by units of time plus existence of non-sampled strata must be evaluated.	WKACCU, 2008	All variables	**Yes** All RDBES CS tables representing sampling stages includes the variable "XXselectionMethod" that allows the distinction between probabilistic and non-probabilistic sampling and methods therein. It is also possible, within the RDBES to compare the temporal distribution of CS data with that of CL and CE data and detect systematic departures from the sampling rates expected for specific time domains.	-	WKACCU seems to confound bias and sampling error, both of which affect estimates and may lead to departures of estimates from the true totals they try to quantify. We have focused on the bias which is frequently considered more severe from a data quality perspective	WGRDBESGOV "Core group" to discuss WKRDB-EST2 suggestions and evaluate possibilities of detecting this source bias An example report was developed to demonstrate how the RDBES can be used to highlight any such discrepancies. See Annex A3.5.
	5	Appropriate time period and spatial coverage: Biological variables change through time and space and in some time periods and areas their determination may be less accurate than in other. There are recommendations from ices groups (e.g., ICES WKMAT) that orient countries during data collection and reduce biases in analyses.	WKACCU, 2008	All variables	**Yes** Information on timing and space of samples is available in RDBES data model that allows investigation into this sources of bias	-	-	-

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	6	Sampling allocation scheme: estimation of the randomness of the sampling. Is sampling pure random with a sampling protocol well followed, or is sampling allocation made on ad hoc or opportunistic observations? A non random sampling is clearly a source of bias which needs to be reported. In the case of length sampling: Random sampling of boxes/trips: This bias, linked to the follow-up of a sampling protocol focuses more on the randomness of both the choice of boxes to sample (always the top box, vs. real random,) and the choice of trips (opportunistic, real random).	WKACCU, 2008	All variables	**Yes** All RDBES CS tables representing sampling stages include variables of type "selectionMethod" that allow the distinction between probabilistic and non-probabilistic sampling and methods therein.	-	-	-
	7	Non sampled strata: Usually, imputation rules exist for non sampled strata, thus this bias will be an evaluation of the appropriateness of the imputation rules used. E.g., Population of vessels: are all vessels included in the population that forms the sampling frame?	WKACCU, 2008	All variables	**No**	**Partially** Imputation of non-sampled strata will be performed within the ICES Transparent Framework (TAF) so will be documented	A suggestion for declaration of out-of-frame non sampled fractions has been put forward by WKRDB-EST2 to the WGRDBESGOV "Core Group". Inclusion of such a feature will allow the identification of (known) parts of the population that are not included in the sampling frame.	WGRDBESGOV "Core Group to discuss the recommendation.

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	8	In general, the PSU is the first level in hierarchy of sampling units, each representing a cluster of fishing trips, hauls within trips, boxes of fish within hauls etc. For the overall raising procedure to be unbiased, the selection of samples at each stage should be random, and the raising factors are derived from the sampling fraction at that stage.	WKPICS, 2011	All variables	**Yes** In contrast with the previous RDB system, the RDBES allows clear identification of sampling levels used by countries in their multi-stage sampling programmes. Furthermore data-submitters can specify for each level of sampling the selection method they used when selecting samples (see "Sample selection methods" in the RDBES documentation)	-	The RDBES makes substantial progress towards identification and documentation of this bias	
	9	Source of information: it is unlikely that one source of information encompasses the statistics of all fisheries, including the temporal, spatial and fishing activity stratification. In all cases, the advantages and limitations of the sources used should provide a clear view on the related bias.	WKACCU, 2008	All variables	**Partially** RDBES documents the source of information of CL and CE data but not the source of information used in defining the sampling frames of CS data.	**No**	Variables that document the sources used to derive the sampling frames of CS data could be added to RDBES but are unlikely add much value to the data.	WGRDBESGOV "Core group" to keep the issue under review.

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
Protocol	10	Quality assurance protocol: Existence and follow-up of a sampling protocol.	WKACCU, 2008	All variables	**No**	**No**	The RDBES does not contain detailed information on all protocols used in everything that is sampled (e.g., age reading, some specifics of onboard or onshore sampling, etc). Supplementary information will be needed to quality assure those aspects To facilitate finding this information the name of the sampling scheme used in the RDBES should be the same as the name used in other sources (e.g. national worplans)	Information on quality assurance protocols should be included in national workplans
	11	Sampling protocol: Existence and adherence to a sampling protocol that yields representative selection of fish for length measurements.	WKACCU, 2008	All variables	**No**	**No**	See comments in issue 10	Information on sampling protocols should be included in national workplans
	12	Non response rate: the percentage of refusal is one of the most important sources of bias for on-board observers. This case discussed in general in Cochran, 1977 has also been addressed by the recent workshop on discards (Anon, 2003) in the frame of the DCR.	WKACCU, 2008	All variables	**Yes** All RDBES CS tables representing sampling stages includes variable "XXsampled" and "XXreasonNotSampled" that allow the recording of different types of non-responses, including refusals.	-	Non-responses are a source of potential bias in the calculation of all parameters	An example report was developed that demonstrates how plots build from RDBES data can highlight the presence of non-responses in the data. See Annex A3.5.

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	13	Responses. WKPICS highlighted the importance of recording non-events, such as documenting failed sampling attempts where procedures were followed but fishermen or merchants barred access to landings or a trip. These events could create bias so need to be accounted for in raised estimates.	WKPICS, 2011	All variables	**Yes** The RDBES allows the recording of non-responses (see "Non-responses and missing values due to quota sampling" in the RDBES documentation)	-	The RDBES will allow this potential source of bias to be analysed.	An example report was developed that demonstrates how plots build from RDBES data can highlight the presence of non-responses in the data See Annex A3.5.
	14	Who collects the samples, the staff responsible for sampling, by the crew or by the port staff? There are potential conflicts of interest in some of these players that may induce bias in sampling selection that then propagates to final estimates.	WKACCU, 2008	All variables	**Yes** All RDBES CS tables representing sampling stages includes variable "sampler" that allows the distinction between different sources of data, including self-sampling, control, observer, etc. That information can be used to evaluate this source of bias.	-	The RDBES will allow this potential source of bias to be documented.	
	15	Species replacement: species thrown away (discarded) because replaced by another. This behaviour, linked to the carriage capacity, must be evaluated if it occurs, either by a well-designed sampling programme (no bias) or by external source (risk of bias).	WKACCU, 2008	All variables	**No**	**No**	Unclear description. The RDBES does not seem to be able to analyse this behavior.	-

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	16	Damaged fish landed: some cases were reported of fishers proposing for sale incomplete, i.e., fish partially cut for any reason, such as bite by a shark.	WKACCU, 2008	Landing and discard weight	**No** RDBES CL table includes variable "CLcatchCategory" that under code "RegDis" allows the reporting of logbook registered discards, exemptions to the landing obligation and damaged fish**.** The current code list does not yet allow for the separation between these categories.	**Yes**	This feature could be added in an upcoming update to the data model. CS data can be reported in different levels of processing. Analyses of these records may provide for additional indications of potential bias in official estimates.	WGRDBESGOV "Core group" to consider possibilities of improving code list and fully detecting this source of bias in the future
	17	Slipping behaviour: In general, this behaviour is linked to specific fisheries such as pelagic trawling. The more or less rare occurrence of rejecting all the catch before it comes on the vessel deck needs to be evaluated.	WKACCU, 2008	Slipping weight	**Yes** RDBES CS data model allows the recording of slipping events as samples associated to variable "SSobservationActivityType"	-	-	-
Quality / precision	18	Working conditions: evaluating the sampled weight with a scale needs proper conditions, which are not always possible. Sampling for discards needs also good conditions for taking the sample and enough time and space for carrying the scientific work. Any constraint on working condi-tions may lead to a bias in the final estimates.	WKACCU, 2008	Landing and discard weight	**Partially** The RDBES BV table allows the declaration of different types of measurement equipment. That information can provide insight into the accuracy of some measurements.	**Yes**	The WGRDBEDGOV "Core group" have been discussing the implementation of a quality scale for all biological measurements.	The WGRDBEDGOV "Core group" are considering the issue

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	19	Staff trained for age reading: information such as the time since the last training or information on the experience of the staff are the elements to determine the risk of bias on age reading. Some international calibration workshop evaluate the competence of age readers for estimating age structure for stock assessment purpose, Age readers formally approved by such a forum, would lead to an absence of bias for this parameter; experience of the staff is an element to determine the risk of bias on estimating the sex of certain species (e.g., Pandalus).	WKACCU, 2008	Age structure, sex-ratio	**No** The RDBES is not capable of holding information of this kind.	**No**	Supplementary information will be needed that can be included in national protocols and/or workplans. This is part of the overall aspects of training and possibilities to increase accuracy of biological determinations	Information on this type training could be requested in national workplans.

	20	Quality documentation on biological variables: Existence of a validity control for the appropriateness of the reading to evaluate the true age (check with tagging or in vivo growing programs).; Existence of a recent age reading calibration workshop; Existence of a recent international exchange in or-der to compare the results of age reading by several readers from different countries on the same material. Usually, the exchange is carried out in preparation of an age reading workshop or at regular interval to assess the need of convening such a workshop. ; International reference set: Existence and routinely use of an agreed inter-national reference set. ; The risk of bias is inherent to the species/stock itself, depending on the difficulty of reading the age. The inter-national calibration workshops use software able to evaluate such a bias; Existence of a routine calibration validation of the equipment used.; How are immature issues being addressed? Is the method used well described and approved? ; Existence and follow-up of an international sampling protocol (ICES WKMAT 2007, survey protocols); existence of a protocol for dealing with immature;		All biological variables	**No** The RDBES is not capable of holding information of this kind. Supplementary information will be needed that can be included in national protocols and/or NWP.	**No**	This is part of the overall aspects related to quality assurance of biological data collection done at national level. The capacity to link between RDBES CS data and the system used in International exchanges and calibrations (SmartDots) is present but MS will need to ensure they use the same individual fish ids when uploading data.	There are on-going discussions between the WGRDBESGOV and the WGSMART data governance groups.

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
		Existence and follow-up of internationally agreed references for histology						
	21	Processing and evaluation methods for age, sex and maturity: Some reading methods are known to be biased for estimating some or all ages.	WKACCU, 2008	Age structure	**Partially** The RDBES BV table includes information on processing method for all biological measurements. Presently only codes for age structures exist, but codes for other relevant methods can be added.	**Yes**	The present RDBES code list would need to be expended.	WGRDBESGOV "Core group" to discuss and evaluate possibilities of detecting this source bias
	22	Staff trained for species identification: information such as the time since the last training or information on the experience at sea are the elements to determine the risk of bias on species identification at the end of a sampling. This source of bias must be combined to the previous one as on one hand a species easy to identify do not present major risk of bias even for a novice, and on the other hand a species difficult to identify is not a source of bias if sampled by a taxonomist.	WKACCU, 2008	All variables	**No** The RDBES is not capable of holding information of this kind. Supplementary information will be needed that can be included in national protocols and/or workplans.	**No**	This is part of the overall aspects of training and possibilities to increase accuracy of taxonomic identifications	Information on this type training could be requested in national workplans.

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	23	Existence of an identification key: photographs or sketches of species of relevance in a given fishing area are very useful tools to ensure correct species identification. The absence of such identification keys, however, is not to be considered a source of bias when the staff that conduct the species identification is trained and experienced in taxonomy.	WKACCU, 2008	All variables	**No** RDBES is not capable of holding information of this kind. Supplementary information will be needed that can be included in national protocols and/or NWP.	**No**	This is part of the overall aspects of training and possibilities to increase accuracy of taxonomic identifications	Information on existence or absence of this type of material should be included in national protocols and workplans
	24	Species subject to confusion: The risk of bias is inherent to the species itself, depending on the difficulty of its identification. A way of evaluating the bias could be through a reference table of species to be agreed by an inter-national forum. The setting of such a table, specific to fishing areas/regions, should be addressed by the ICES PGCCDBS.	WKACCU, 2008	All variables	**Partially** The RDBES is not capable of knowing whether a recorded species has been mis-identified. The RDBES SL table contains the taxa recorded in each sampling event. The table can be compared with regional standards for completeness. If MS provide that level of detail, differences between SL tables of different sampling events may indicate lack of consistency between observers		Reference tables of species are now defined in EU legislation on data collection. However, not all sampling programmes in each country sample all the species in that list (some are a subset of species or are species focused)	Information on the species targeted by each sampling programme should be requested in national workplans. Those target species should be recorded in the SL table of RDBES

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	25	Unit definition: Existence and follow-up of an international agreed definition and specifications. Effort statistics obtained through a census or a sampling programme.	WKACCU, 2008		**Yes** RDBES CE table includes variables "CEdataTypeForScientificEffort" and "CEdataSourceForScientificEffort" that provide that information. RDBES CE also requests effort estimation to be carried out according to STECF guidelines (WKTRANSVERSAL II report)	-	-	-
Analysis	26	Size of the catch effect: When catches are big and only a guestimated fraction has been sampled, the bias is more likely than when a significant fraction of the catch (say more than 10%) is taken for sampling. In general this information is absent even from the raw samples.	WKACCU, 2008		**Partially** The RDBES allows the quantification of sampling fractions in all CS tables representing sampling stages The RDBES can store the number of units sampled and the total number of units at each stage – it is then possible to identify small samples taken from large hauls. It is not currently possible to record how, for example, the size of a haul was calculated (e.g. measured or estimated).	-	-	WGRDBESGOV "Core group" to discuss and evaluate possibilities of detecting this source bias
	27	Changes in fishers' behaviour when observed, on board sampling.	WKPICS, 2011	All biological variables	**No** The RDBES can't identify this source of bias.	**No**	-	-

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	28	Taxonomic changes: changes in species nomenclature over time, e.g. the splitting of sandeel species in the face of new knowledge, may impact the consistency of a time-series.	WKACCU, 2008	All variables	**Partially** New uploads to the RDBES will be required to use the latest, valid aphia id for a species. An aphia id may become invalid – in this case MS can be asked to re-upload data using the new code, or the RDBES host can update the aphia id. In RDBES consistency of species recording in time can be checked by comparing SL records across time periods.	-	-	Information on the species targeted by each sampling programme should be requested in national workplans. Those target species should be recorded in the SL table of RDBES
	29	Area misreporting: like for the species misreporting, there may be a sudden increase of a species reported in an uncommon neighboring area. This type of bias may be assessed by checking the consistency between different sources e.g. logbooks, VMS, sales notes, questionnair surveys, cpue trends of commercial vs. surveys, …	WKACCU, 2008	Landings weight, effort	Partially Not all different sources of information will be incorporated in the RDBES. But RDBES CS data can be used to estimate landings of a species in specific spatial domains that can then be evaluated against similar estimates in RDBES CL data and/or compared with other types of external data like VMS, survey data, etc.	-	WGRDBESGOV is promoting workshops (e.g., WKRDB-EST) with the aim of developing code for the RDBES / TAF. Some of that code will allow the identification of situations where this type of bias may impact final estimates.	-

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	30	Missing part: ratio between the retained fractions estimated on-board by observers and the landings of a species. A statistical test can be performed to evaluate if the slope is significantly different from one.	WKACCU, 2008	Landing weight	**Partially** RDBES CS data can be used to estimate total landings of a species that can then be compared with RDBES CL data	-	See the comment on issue 29	-
	31	Quantity misreporting: known as the most current bias in fisheries statistics, this bias may be assessed together with area misreporting and with the addition of sources like economic surveys and EU control database.	WKACCU, 2008	Landings weight, effort	**Partially** RDBES CS data can be used to estimate landings of a species in specific spatial domains that can then be evaluated against similar estimates in RDBES CL data and/or compared with other official data	-	See the comment on issue 29	-
	32	Species misreporting: A sudden increase of an unexpected species may occur in the statistics, thus pointing out a potential risk of species misreporting. This case is generally linked to quota consumption. Another way of detecting such a bias is dissimilarities between on-board observers reporting for the same fishing activity, or dissimilarities between on-board observers and landing statistics	WKACCU, 2008	All variables	**Partially** RDBES CL data can be analysed to identify changes in species reporting over time. Further, Estimates derived from RDBES CS data may be compared with the official estimates provided in CL table to identify this type of bias.	-	See the comment on issue 29	-

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	33	Management measures leading to discarding behavior: the specification of the measure and the date of entry into force are indications of potential bias, if not monitored through a well-designed sampling program.	WKACCU, 2008		**No** The time resolution of RDBES CS data should allow for this type of analysis but management measures impacting the estimates are multiple and have different types of impacts thus not being possible to document with the data.	**No**	-	Information on impacts of management measures can be pursued by independent projects looking into this issue.
	34	High grading; selecting a given size range for landing a species depending on the market demand or to reduce the quota consumption automatically change the discarding ogive. High grading behaviour may be evaluated by interviews and/or on-board observers.	WKACCU, 2008		**Partially** RDBES CS data from market sampling can be compared to onboard sampling to evaluate high-grading in commercial landings	-	-	-
	35	Change in selectivity: bias linked to the characteristics of the gear and evaluation whether the length structure sampled is representative of the exact characteristics of the gears used at the population level.	WKACCU, 2008		**Yes** RDBES CS data can be used to investigate these type of biases	-	-	-

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	36	Statistical processing: when direct biological measurements (e.g., age reading, mean weight) are impossible, statistical methods may be used to estimate those variables that may introduce bias in analysis (e.g., length-weight relationships, Von Bertalanffy models). The time between the references used for modelling and the actual time strata is an indication on the potential induced bias.	WKACCU, 2008	Age structure, mean-weight	**Partially** In principle RDBES CS data will allow these types of analyses. Estimates carried out under RDBES will be registered in TAF allowing future discussion on appropriateness of methods and calculations used. Length-weight relationships, von Bertalanffy and many other types of models can be derived from RDBES data and used to evaluate the biases incurred when using the different methods.	-	-	-
Analysis / Estimation	37	Incomplete ALK: Appropriateness of the imputation rules used, e.g., for filling length classes without age information.	WKACCU, 2008	Age structure	**Partially** Estimates carried out under RDBES will be registered in TAF allowing future discussion on appropriateness of methods and calculations used.	-	-	-

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	38	Specific handling of some biological variables to reduce bias (e.g., Plus group: bias linked to the setting of the plus group, and the existence or not of international agreement; Skipped spawning: following ICES WKMAT recommendation, is skipped spawning known to happen and taken into account?; Catchability effect: for some species the catchability by sex vary over time. If such behaviour related change in catchability occurs, do the estimates take this into account following an agreed protocol?)	WKACCU, 2008	All variables	**Partially** Estimates carried out under RDBES will be registered in TAF allowing future discussion on appropriateness of methods and calculations used.	-	-	-
	39	Raising variable: For raising to the population, different raising procedures must be compared and also raising the retained fraction to be compared with the landing statistics is a solution to assess the relevance of the variable used for raising (WKDRP, 2007).	WKACCU, 2008	All variables	**Partially** RDBES is designed in a way that makes it possible to estimate catches by a variety of probabilistic methods including ratio estimation. The specifics of ratio estimation within the RDBES have not yet been developed.	**Yes**	There are initiatives are planned that make these analyses possible (e.g. WKRATIO scheduled for 2021))	WKRATIO to provide indications on analyses comparing different methods. WGRDBESGOV to include implementation of those analysis in the system.

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
	40	Conversion factor: information such as the age and the methodology used for the conversion factor, are indications on the potential induced bias. The magnitude of the conversion factor used is also an indication, e.g. estimating landing weight from fillet or from gutted fish will lead to different amplification of a bias.	WKACCU, 2008	landings and discrad weight, length frequencies, mean weights	**Partially** RDBES SA table includes variable "SAconversionFactorMesLive" that allows preliminary investigations into this source of bias (e.g., identification of conversion factors that changed through time, different conversion factors between countries). Full information on the year, methodology and/or source for the conversion factor is not presently available in the data model and should be and included in national protocols,	**Yes**	Inclusion of similar conversion factors in BV may be needed	WGRDBESGOV "Core group" to evaluate possibilities of fully detecting this source of bias and help to improve conversion factors

Issue Category	Id	Issue description	Source	Affects	Can the RDBES tackle it at the moment?	Could the RDBES potentially tackle this in the future?	Comment	Actions
Estimation	41	Grouping statistics: some commercial naming include groups of several species, e.g. lophius, megrims. It may also be the case that a commercial naming includes incidentally other species, as often encountered with the elasmobranchs (e.g. mixture of ray species in a box of Raja clavata). Scien-tific sampling surveys are generally used to quantify the percentage of species within the relevant commercial names, and if it is the case, there is no major risk of bias.	WKACCU, 2008	All variables	**Yes** The RDBES SL table includes both commercial and scientific denominations allowing for the recording of sampled species composition within commercial denominations. Those recordings can be used to split national statistics of very aggregated categories (e.g., split national landings of monkfishes into landings of angler and black-bellied angler)	-	This feature has been implemented but not fully tested with real data yet.	WGRDBESGOV to complete tests of this feature within the WKRDB-EST framework

Annex A3.5 Example reports to illustrate potential sources of bias

Comparison of sample data to population data report

The following plots illustrate graphical reports made using the RDBES data format that can be used by MS to inspect their data and help them identify errors in their datasets and sources of potential bias in their catch estimates. The R code for production of these plots can be found on Github[17] and had as a starting block earlier developments achieved during 2019 RCG intersessional subgroups on fisheries and sampling overviews.

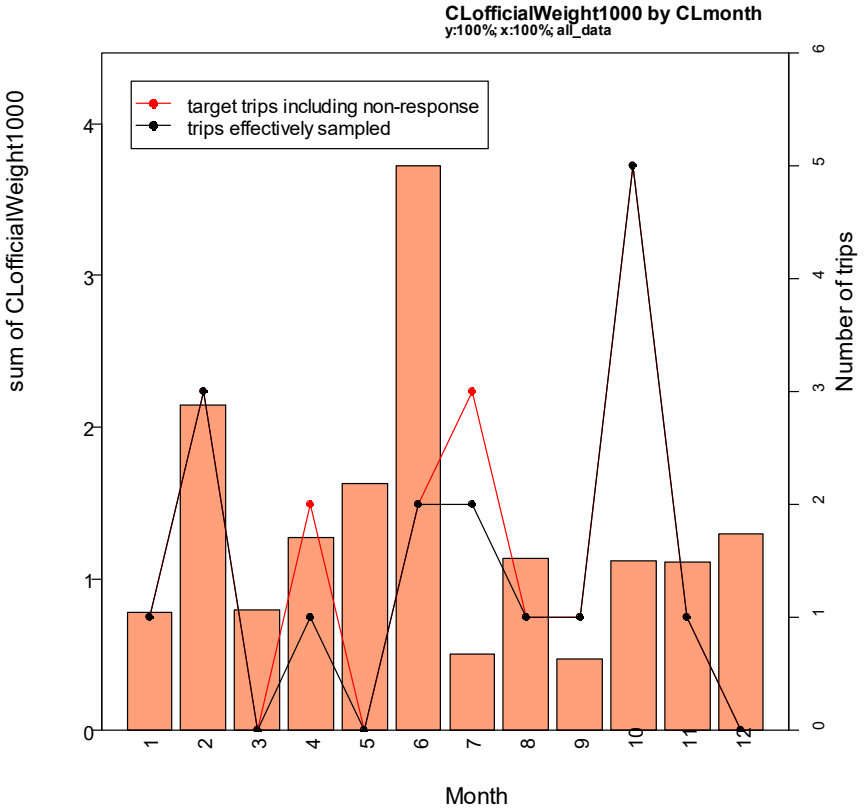
Note: These type of plots should be used with caution since the patterns they may highlight can only be evaluated with full knowledge of the original sampling plan and its implementation (i.e. by combining RDBES, national sampling plan information, and implementation notes) and some observed patterns might be explained by either biases in sampling or by the natural variation occurring in a probabilistic sampling scheme.

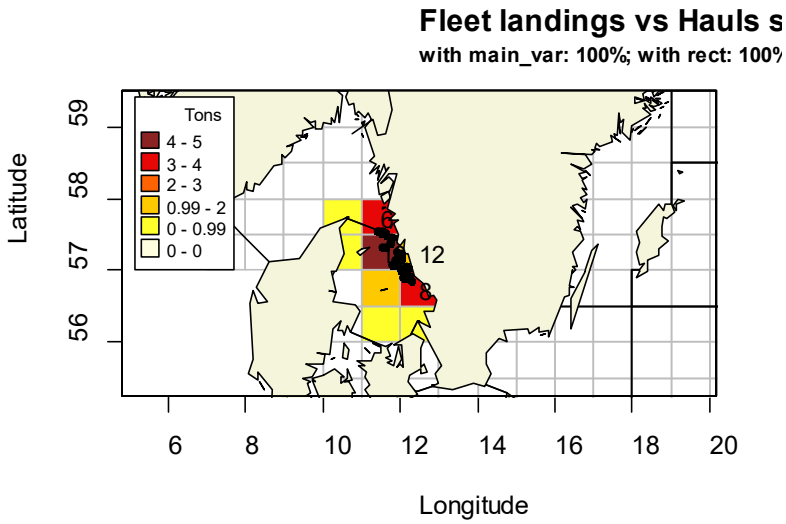
The first figure displays landings per month from a certain fleet (in tonnes, bars, left y-axis) alongside numbers of trips sampled including and excluding non-response (a potential source of bias). By analysing plots like this, data submitters will be able to scrutinize aspects of quality such as errors in data (e.g. is the large number of samples in month 10 real or an error in the data submitted?). In parallel, the graph provides data estimators with a quick overview of the data available along a temporal dimension, facilitating the consideration of potential sources of bias such as non-responses (e.g. could the non-responses observed impact the estimates?) and departures from the sampling plan during its implementation that may need to be accounted for during estimation to avoid biases (e.g. if the goal of this plan was to sample 2 trips per month, clear departures in the implementation occurred in many months). Additionally, the graph may also allow insight into potential precision issues such as those brought about by lack of coverage in months where landings could be known to be important and highly variable.

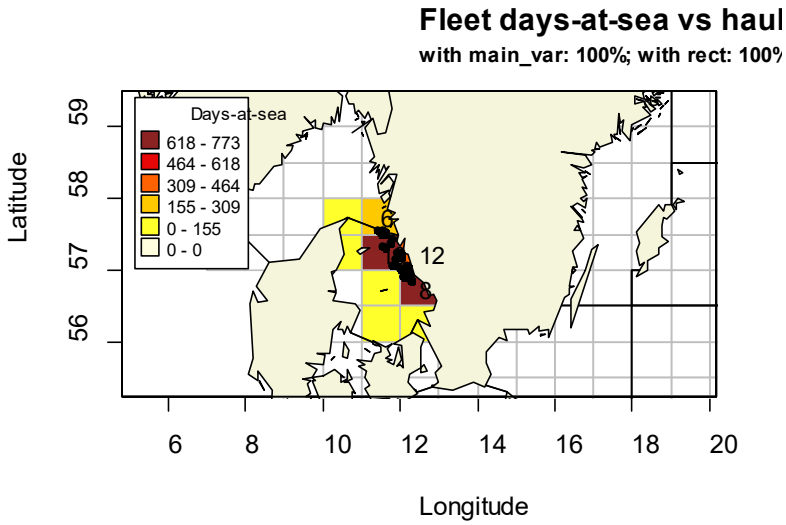
The second figure displays similar data (population and sampled totals) but focuses on the spatial coverage. The figure displays landings (right map, in tonnes) and effort (left map, as days-at-sea) per ICES statistical rectangle alongside the number of hauls samples and their position. These type of plot can be used to visually check for potential biases in the sampling data such as those caused by "observer effects", but is also useful to identify biases potentially caused by errors in the data (e.g., a sample collected where no effort took place) or to visually check for abnormally high concentration of data in some regions and not others (either probabilistic, or more importantly, non-probabilistic).

Different variations of both these type of plots can be coded, e.g., in the case of the bar plot one might be interested in plotting effort instead of landings or seeing the data at different time resolutions (e.g. quarters instead of months). It most likely that several of these graphs will need to be combined during analyses of potential biases (e.g. if sampling goals are established at quarterly level, data estimators may still be interested in monthly bar plots to check if implementation issues such as all trips being sampled in the first month of each quarter could potentially impact the estimates).

[17] https://github.com/ices-eg/WK_RDBES/tree/master/Special_Request_20_05


CLofficialWeight1000 by CLmonth
y:100%; x:100%; all_data
target trips including non-response
trips effectively sampled
sum of CLofficialWeight1000
Number of trips
Month
0
1
2
3
4
5
6
7
8
9
10
11
12





Summary of selection methods report

This report provides a summary of the different selection methods used in a sampling scheme – in particular it highlights the number of units that were selected using probabilistic and non-probabilistic methods at each stage. In the case where the data contains information on un-sampled units these are also summarized. The R code can be on Github[18].

[18] https://github.com/ices-eg/WK_RDBES/tree/master/Special_Request_20_05

Bias - Summary of selection methods and declaration of none sampled, ESP-AZTI_DCF_Onboard_Sampling 1966

Selection methods

The selection methods are grouped the following way

Code	Description	Selection method type
CENSUS	Census (CENSUS)	Probabilistic
NotApp	NotApp	NA
NotSam	NotSam	NA
NPAH	Ad Hoc Sampling (NPAH)	None-probabilistic
NPEJ	Expert Judgement (NPEJ)	None-probabilistic
NPQS	Quota Sampling (NPQS)	None-probabilistic
SRSWOR	Simple Random Sampling Without Replacement (SRSWOR)	Probabilistic
SRSWR	Simple Random Sampling With Replacement (SRSWR)	Probabilistic
SYSS	Systematic Sampling (SYSS)	Probabilistic
UPSWOR	Unequal Probability Sampling Without Replacement (UPSWOR)	Probabilistic
UPSWR	Unequal Probability Sampling With Replacement (UPSWR)	Probabilistic

NotApp and NotSam are not considered in these overviews.

Overall - summary

Number of samples selected in a probabilistic and none-probabilistic way per sampling level.

The use of probabilist selection is highlighted in green and number of none-probabilistic selections are highlighted in red.

Sampling scheme	Year	Country	Sampling unit level	Sampling unit type	Number of probabilistic selections	Number of none-probabilistic selections
ESP-AZTI_DCF_Onboard_Sampling	1966	DK	1	Vessel	10	10
ESP-AZTI_DCF_Onboard_Sampling	1966	DK	2	Fishing trip	12	8
ESP-AZTI_DCF_Onboard_Sampling	1966	DK	3	Fishing operation	60	0
ESP-AZTI_DCF_Onboard_Sampling	1966	DK	4	Species selection	60	0
ESP-AZTI_DCF_Onboard_Sampling	1966	DK	5	Sample	120	0

Detailed summary per sampling level

Number per selection methods used per sampling level.

In each table the stratum from the level above is stated to make identification easier.

1. sampling unit

Sampling unit type	Sampling scheme	Year	Country	Stratum above	Stratum here	Selection method	Number of samples
Vessel	ESP-AZTI_DCF_Onboard_Sampling	1966	DK	DE_stratum1	VS_stratum1	NPEJ	5
Vessel	ESP-AZTI_DCF_Onboard_Sampling	1966	DK	DE_stratum1	VS_stratum2	SRSWR	5
Vessel	ESP-AZTI_DCF_Onboard_Sampling	1966	DK	DE_stratum2	VS_stratum1	NPEJ	5
Vessel	ESP-AZTI_DCF_Onboard_Sampling	1966	DK	DE_stratum2	VS_stratum2	SRSWR	5

2. sampling unit

Sampling unit type	Stratum above	Stratum here	Selection method	Number of samples
Fishing trip	VS_stratum1	U	NPAH	3
Fishing trip	VS_stratum1	U	SRSWR	7
Fishing trip	VS_stratum2	U	NPAH	5
Fishing trip	VS_stratum2	U	SRSWR	5

3. sampling unit

Sampling unit type	Stratum above	Stratum here	Selection method	Number of samples
Fishing operation	U	U	SRSWOR	60

4. sampling unit

Sampling unit type	Stratum above	Stratum here	Selection method	Number of samples
Species selection	U	U	CENSUS	60

5. sampling unit

Sampling unit type	Stratum above	Stratum here	Selection method	Number of samples
Sample	U	U	SRSWOR	120

None sampled

Overall - summary

Only sampled data declared

Sampling scheme	Year	Country	Sampling unit level	Sampling unit type	Sampled or not	Industrial Decline	No Answer	No Contact Details	Not Avalible	Observer Decline	Other	Quota Reached	No reason given

Detailed summary per sampling level

Number of none sampled per reason for not sampling.

In each table the stratum from the level above is stated to make identification easier.

1. sampling unit

Only sampled data declared

Sampling unit type	Sampling scheme	Year	Country	Stratum above	Stratum here	Sampled or not	Reason for not sampling	Number of samples

2. sampling unit

Only sampled data declared

Sampling unit type	Stratum above	Stratum here	Sampled or not	Reason for not sampling	Number of samples

3. sampling unit

Only sampled data declared

Sampling unit type	Stratum above	Stratum here	Sampled or not	Reason for not sampling	Number of samples

4. sampling unit

Only sampled data declared

Sampling unit type	Stratum above	Stratum here	Sampled or not	Reason for not sampling	Number of samples

Annex 4: Resolution

The **Second Workshop on Estimation with the RDBES data model (WKRDB-EST2)** chaired by Nuno Prista, Sweden and Kirsten Birch Håkansson, Denmark, will meet through a web meeting from 14 to 18 September 2020 to:

a. Development and documentation R scripts for design based estimation for each hierarchy in the RDBES data model (supporting Advice Plan: Assuring Quality);
b. Identify and document issues problems with RDBES data model relating to design based estimation (supporting Advice Plan: Assuring Quality);
c. Develop roadmap for future improvements to the estimation procedures within the RDBES;

WKRDB-EST2 will present a written report to ACOM by 18 December 2020.

Supporting information

Priority	This workshop is considered of very high priority. The activities of this workshop will promote the development of a Regional Database and Estimation System (RDBES) by developing the algorithms and code required for design based estimation within the upcoming RDBES. The RDBES will be integrated in TAF and work as a database for both ICES and the Baltic Sea, North Sea & Eastern Arctic, and North Atlantic Regional Coordination Groups (RCGs), producing high-quality, transparent, estimates for ICES Fisheries Advice.
Scientific justification	Term of Reference a) The R-scripts started at WKRDB-EST in 2019 will be further developed towards full implementation of design-based estimation and the production of point estimates of fisheries variables such as catch volumes, numbers-at-length and number-at-age. Development will be based on countries data from the different hierarchies uploaded during the upcoming RDBES data call (September 2019) and extracted from the system prior to the meeting. The R-code will be documented with associated statistical formulas and used in RDBES documentation. The development of scripts for other estimation methods (e.g., ALK-based estimation, Ratio-Estimation) will not be addressed during the WK but aspects like post-stratification and domain estimation will be included in the code if time allows. Term of Reference b) The development of R scripts for design-based estimation based on the RDBES data model is an important test point within the development of the RDBES. Issues identified during the WK that limit the application of design-based estimation in the RDBES will be documented and forwarded to the RDBES development group for further discussion. Term of Reference c) Design-based estimation is not the only type of estimation used to produce commercial catch estimates within the ICES community. Model-assisted and model-based estimation are two commonly used alternatives that require theoretical and code development in the context of RDBES and that are being explored by other EGs (e.g., WGCATCH). At the end of WKRDB-EST, based on the progress they have achieved in design-based estimation during the week, WK participants will jointly reflect on the best way forward for further development of RDBES estimation routines. Both SCRDB and WGCATCH will be informed on the conclusions of these discussions.
Resource requirements	The two co-chairs and the rest of the active members of the core group of RDBES development will be requested to participate and coordinate algoritm and code development ahead of the meeting.

Participants	Max 20 people. Participants should be proficient in writing own scripts and functions in R language and/or have good knowledge of survey sampling and estimation.
Secretariat facilities	None.
Financial	None.
Linkages to advisory committees	There are no direct linkages with the advisory committees, but there is a direct link to SCRDB and close links to activities of WGCATCH, WGBIOP, WGBYC, and PGDATA. Stock assessment Working Groups will ultimately use and benefit from quality estimates produced within the RDBES.
Linkages to other committees or groups	
Linkages to other organizations	The RDBES estimates are connected to regional data collection defined by the RCGs under the European Commission, EC. The RDBES will also support the ICES countries in providing data for assessment. In the case of EU MS, the RDBES is expected to facilitate and improve the quality of provision of commercial catch data requested under different data calls.

Annex 5: Design-based estimation for a three stage sampling design

Sampling without replacement in all three stages

Consider the following sampling design in three stages where the primary sampling units are vessels, the secondary sampling units are trips and the tertiary sampling units are hauls.

Stage I: Sampling of vessels

A random sample without replacement of vessels is drawn from all the vessels in the population. The set of vessels in the population is denoted U_I of size N_I and the sample of vessels is denoted s_I of size n_I. Each vessel is looked upon as a cluster of trips.

Stage II: Sampling of trips

For every vessel i selected in stage I, a random sample without replacement of trips is drawn from all the trips associated with the vessel. The set of trips associated with vessel i is denoted U_{IIi} of size N_{IIi} and the sample of trips is denoted s_{IIi} of size n_{IIi}. Each trip is looked upon as a cluster of hauls.

Stage III: Sampling of hauls

For every trip q selected in stage II, a random sample without replacement of hauls is drawn from all the hauls associated with the trip. The set of hauls associated with trip q is denoted U_{iq} of size N_{iq} and the sample of hauls is denoted s_{iq} of size n_{iq}.

For each haul k selected in stage III, the weight of discards, y_k, is observed. The problem is to estimate the total weight of discards for all possible hauls, trips and vessels,

$$t_y = \sum_{U_I} \sum_{U_{IIi}} \sum_{U_{iq}} y_k$$

and the variance of this estimator. To accomplish this, we need the inclusion probabilities for each stage.

1.1 Inclusion probabilities for the general case

For **stage I**, the first order inclusion probability π_{Ii} is the probability of vessel i to be included in the sample s_I. The second order inclusion probability π_{Iij} is the joint probability of vessel i and j to be included in s_I.

For **stage II**, the first order inclusion probability $\pi_{IIq|i}$ is the conditional probability of trip q to be included in the sample s_{IIi} (conditional on the stage I sampling). The second order inclusion probability $\pi_{IIqr|i}$ is the conditional joint probability of trip q and r to be included in s_{IIi}.

For **stage III**, the first order inclusion probability $\pi_{k|iq}$ is the conditional probability of haul k to be included in the sample s_{iq} (conditional on the stage I and II sampling). The second order inclusion probability $\pi_{kl|iq}$ is the conditional joint probability of haul k and l to be included in s_{iq}.

We summarize these general inclusion probabilities in the table below.

Stage	Inclusion probabilities, general	
	First order	Second order
I	π_{Ii}	π_{Iij}
II	$\pi_{IIq\|i}$	$\pi_{IIqr\|i}$
III	$\pi_{k\|iq}$	$\pi_{kl\|iq}$

(Note that $\pi_{Iii} = \pi_{Ii}$; $\pi_{IIqq|i} = \pi_{IIq|i}$; $\pi_{kk|iq} = \pi_{k|iq}$.)

1.2 Estimation for the general case

In general, the HT estimator of t_y with respect to all three stages is given by

$$\hat{t}_y = \sum_{s_I} \frac{1}{\pi_{Ii}} \sum_{s_{IIi}} \frac{1}{\pi_{IIq|i}} \sum_{s_{iq}} \frac{y_k}{\pi_{k|iq}}$$

We can also write $\hat{t}_y$ as

$$\hat{t}_y = \sum_{s_I} \frac{\hat{t}_i}{\pi_{Ii}}$$

where $\hat{t}_i$ is the HT estimator of the total weight of discards for vessel i with respect to stage II and III:

$$\hat{t}_i = \sum_{s_{IIi}} \frac{1}{\pi_{IIq|i}} \sum_{s_{iq}} \frac{y_k}{\pi_{k|iq}}$$

Similarly, the estimator $\hat{t}_i$ can be written as

$$\hat{t}_i = \sum_{s_{IIi}} \frac{\hat{t}_{iq}}{\pi_{IIq|i}}$$

where $\hat{t}_{iq}$ is the HT estimator of the total weight of discards for trip q with respect to stage III:

$$\hat{t}_{iq} = \sum_{s_{iq}} \frac{y_k}{\pi_{k|iq}}$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = \sum\sum_{s_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}$$

where

$$\hat{V}_i = \sum\sum_{s_{iq}} \frac{\pi_{kl|iq} - \pi_{k|iq}\pi_{l|iq}}{\pi_{kl|iq}} \frac{y_k}{\pi_{k|iq}} \frac{y_l}{\pi_{l|iq}}$$

Note that for the point estimator we only use the first order inclusion probabilities. For the variance estimator we also need the second order inclusion probabilities.

1.3 Inclusion probabilities for SRS without replacement in each stage

The inclusion probabilities valid for the case of SRS without replacement in each stage are given in the table below.

Stage	Inclusion probabilities, SRS without replacement	
	First order	Second order
I	$\frac{n_I}{N_I}$	$\frac{n_I(n_I-1)}{N_I(N_I-1)}$
II	$\frac{n_{IIi}}{N_{IIi}}$	$\frac{n_{IIi}(n_{IIi}-1)}{N_{IIi}(N_{IIi}-1)}$
III	$\frac{n_{iq}}{N_{iq}}$	$\frac{n_{iq}(n_{iq}-1)}{N_{iq}(N_{iq}-1)}$

1.4 Estimation for SRS without replacement in each stage

For SRS without replacement in each stage, the HT estimator of t_y simplifies into

$$\hat{t}_y = \sum_{s_I} \frac{N_I}{n_I} \sum_{s_{IIi}} \frac{N_{IIi}}{n_{IIi}} \sum_{s_{iq}} \frac{N_{iq}}{n_{iq}} y_k = \frac{N_I}{n_I} \sum_{s_I} \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} \frac{N_{iq}}{n_{iq}} \sum_{s_{iq}} y_k$$

The estimator can also be written as

$$\hat{t}_y = \frac{N_I}{n_I} \sum_{s_I} \hat{t}_i$$

where

$$\hat{t}_i = \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} \frac{N_{iq}}{n_{iq}} \sum_{s_{iq}} y_k = \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} \hat{t}_{iq}$$

and

$$\hat{t}_{iq} = \frac{N_{iq}}{n_{iq}} \sum_{s_{iq}} y_k$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = N_I^2 \frac{1 - n_I/N_I}{n_I} S^2_{\hat{t}s_I} + \frac{N_I}{n_I} \sum_{s_I} \left[N_{IIi}^2 \frac{1 - n_{IIi}/N_{IIi}}{n_{IIi}} S^2_{\hat{t}s_{IIi}} + \frac{N_{IIi}}{n_{IIi}} \sum_{s_{IIi}} N_{iq}^2 \frac{1 - n_{iq}/N_{iq}}{n_{iq}} S^2_{ys_{iq}} \right]$$

where

$$S^2_{\hat{t}s_I} = \frac{1}{n_I - 1} \sum_{s_I} \left[\hat{t}_i - \left(\sum_{s_I} \hat{t}_i / n_I \right) \right]^2;$$

$$S^2_{\hat{t}s_{IIi}} = \frac{1}{n_{IIi} - 1} \sum_{s_{IIi}} \left[\hat{t}_{iq} - \left(\sum_{s_{IIi}} \hat{t}_{iq} / n_{IIi} \right) \right]^2;$$

$$S^2_{ys_{iq}} = \frac{1}{n_{iq} - 1} \sum_{s_{iq}} \left[y_k - \left(\sum_{s_{iq}} y_k / n_{iq} \right) \right]^2$$

1.5 Simplified variance estimation

Some simplified variance estimators for multistage sampling are discussed in Särndal et al (1992, sec 4.6). One possibility is to use only the first term in the expression for the variance estimator; that is, using the abridged HT estimator

$$\hat{V}(\hat{t}_y) = \sum\sum_{s_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Iij}}{\pi_{Iij}} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}}$$

Under SRS without replacement in all stages, this would mean using

$$\hat{V}(\hat{t}_y) = N_I^2 \frac{1 - n_I/N_I}{n_I} S^2_{\hat{t}s_I}$$

If the sample size in stage I is fixed, an alternative is to use the abridged Yates-Grundy estimator

$$\hat{V}^*(\hat{t}_y) = -\frac{1}{2}\sum\sum_{s_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Iij}}{\pi_{Iij}} \left(\frac{\hat{t}_i}{\pi_{Ii}} - \frac{\hat{t}_j}{\pi_{Ij}}\right)^2$$

In both cases, this would lead to underestimation of the true variance. However, if the variance contributions from stage II and III are small, this underestimation might not be so important.

Another option is to do the variance estimation as if vessels were selected with replacement in stage I. The estimation formula for this situation is given in the next section. This approach might in general lead to both over- and underestimation of the true variance.

2 Sampling with replacement in the first stage

Consider again a sampling design in three stages where the primary sampling units are vessels, the secondary sampling units are trips and the tertiary sampling units are hauls. The difference from the design in section 1 is that the sampling is done with replacement in the first stage whereas the sampling in subsequent stages is still without replacement.

Stage I: Sampling of vessels

A random sample *with replacement* of vessels is drawn from all the vessels in the population in such a way that, at every draw, p_i is the probability of selecting vessel i. The set of vessels in the population is denoted U_I of size N_I. The ordered sample of vessels is denoted $os_I = (i_1, \dots, i_v, \dots, i_{m_I})$, where i_v is the vessel selected in draw number v and m_I is the number of draws. Each vessel is looked upon as a cluster of trips.

Stage II: Sampling of trips

For every vessel drawing i_v in stage I, a random sample without replacement of trips is drawn from all the trips associated with the vessel. The set of trips associated with vessel drawing i_v is denoted U_{IIi_v} of size N_{IIi_v} and the sample of trips is denoted s_{IIi_v} of size n_{IIi_v}.

Stage III: Sampling of hauls

For every trip q selected in stage II, a random sample without replacement of hauls is drawn from all the hauls associated with the trip. The set of hauls associated with trip $q \in s_{IIi_v}$ is denoted U_{i_vq} of size N_{i_vq} and the sample of hauls is denoted s_{i_vq} of size n_{i_vq}.

We assume that the sampling in stage II and III has the properties of invariance and independency.

2.1 Estimation for the general case

In general, the HH estimator of t_y with respect to all three stages is given by

$$\hat{t}_y = \frac{1}{m_I}\sum_{v=1}^{m_I}\frac{\hat{t}_{i_v}}{p_{i_v}}$$

where $\hat{t}_{i_v}$ is the HT estimator of the total weight of discards for vessel drawing i_v with respect to stage II and III:

$$\hat{t}_{i_v} = \sum\nolimits_{s_{IIi_v}} \frac{1}{\pi_{IIq|i_v}} \sum\nolimits_{s_{i_vq}} \frac{y_k}{\pi_{k|i_vq}}$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = \frac{1}{m_I(m_I-1)}\sum_{v=1}^{m_I}\left(\frac{\hat{t}_{i_v}}{p_{i_v}} - \hat{t}_y\right)^2$$

(see Särndal et al, 1992, Result 4.5.1).

2.2 Estimation for SRS with replacement in the first stage

For SRS with replacement in stage I, the drawing probability p_{i_v} is equal to $1/N_I$ for all vessel drawings i_v. If SRS without replacement is used in stage II and III, the HH estimator of t_y simplifies into

$$\hat{t}_y = \frac{1}{m_I}\sum_{v=1}^{m_I}\frac{\hat{t}_{i_v}}{p_{i_v}} = \frac{N_I}{m_I}\sum_{v=1}^{m_I}\hat{t}_{i_v}$$

where

$$\hat{t}_{i_v} = \frac{N_{IIi_v}}{n_{IIi_v}} \sum\nolimits_{s_{IIi_v}} \frac{N_{i_vq}}{n_{i_vq}} \sum\nolimits_{s_{i_vq}} y_k$$

An unbiased estimator of the variance of $\hat{t}_y$ is given by

$$\hat{V}(\hat{t}_y) = \frac{N_I^2}{m_I(m_I-1)}\sum_{v=1}^{m_I}\left(\hat{t}_{i_v} - \hat{\bar{y}}_U\right)^2$$

where $\hat{\bar{y}}_U = \hat{t}_y/N_I$.

Reference

Särndal, C.-E., Swensson, B., Wretman, J. (1992) Model Assisted Survey Sampling. Springer-Verlag.

Annex 6: Notes on pending issues in the area of design-based estimation using the RDBES data model

Note: the solutions presented in this annex do not represent definitive conclusions of WKRDB-EST2 on any of the issues, but rather reflections and suggestions of the EG that should be further considered by the core-group of development of RDBES and, if necessary, by a future edition of the present WK.

Issue #1 Declaring out-of-frame components of the study population

Present situation

Guidelines in documentation (v1.18) state that all strata included in the sampling frame of a particular sampling stage must be reported even if they have not been sampled. When a stratum has not been sampled it is reported as a row where selection method = "not-sampled" and sample size = 0. With regards to parts of the target population of each sampling stage that are not covered by the study population and therefore are not in the sampling frame (a.k.a. "out-of-frame"), the documentation states that they should not be declared. Particular attention is given to the SA table where stratification of species landings in size categories is frequently reported. Here documentation is more explicit - if variables such as SAcommSizeCat are present, it is still important that the full stratification is declared in the stratum column as this will be the one considered for effects of partitioning the population during the estimation and that in those cases all size categories present (e.g., in a landing) should be reported in the sample table even if they were not sampled but size categories absent from that landing need not be reported (i.e., they are considered in-frame). Examples are provided that inform on how to report missing data in both VS and SA tables. Some situations exist where national programmes target specific size categories (i.e., aim to sample some size categories but not others) but these are not addressed in documentation.

Details of Issue

Difficulties related to the non-reporting the "out-of-frame" component(s) of the target population become apparent in Example 1, extracted from the documentation (see below). In that example, that follows v1.18 guidelines, all the "in-frame" strata had been sampled and declared. A data user might suspect the existence of some vessels in a VSstratum <10m were not considered in the sampling frame but cannot be sure of that situation. Even if he would know about the existence of <10m vessels in the fleet, he would not have available the size of that unsampled fleet relative to the fleet sampled. Accordingly, he would not have the necessary elements to evaluate the importance (or not) of accounting for those smaller vessels during calculation of population-level estimates. Example 2 in documentation, from which example 3 was derived, reveals that 1000 vessels were in VSnumberTotal, i.e., a larger number than all vessels >=10 m combined.

Example 1 (present situation): Simplified example of stratification by vessel size where one stratum is out-of-frame (<10 m). In gray a stratum that does not presently need to be declared according to RDBES documentation. Following current guidelines, the existence of these additional "out-of-frame" vessels should not be declared in RDBES.

VSid	VSstratification	VSstratum	VSselectionMethod	VSnumberTotal	VSnumberSampled
1	Y	>=10 <15m	random	500	2
2	Y	>=10 <15m	random	500	2
3	Y	>=15m	random	200	2
4	Y	>=15m	random	200	2
---	---	<10m	----	1000	0

Other examples exemplifying this issue are displayed in Example 2, 3 and 4.

Example 2 (present situation): Simplified example of stratification hauls in a trip by subdivision. The sampling programme targets 27.3.a.20 but not 27.4.a. n = 6 hauls in 27.4.a (in gray) took place during the observed trip. These hauls are not in the sampling frame of the programme. Following current guidelines, the existence of these hauls should not be declared in RDBES. Without clearly flagging such situation estimates like total landings or discards in this trip would be severely biased.

FOid	FOstratification	FOstratum	FOselectionMethod	FOnumberTotal	FOnumberSampled
1	Y	27.3.a.20	random	4	2
2	Y	27.3.a.20	random	4	2
---	---	27.4.a	----	6	0

Example 3 (present situation): Simplified example of stratification of a landing event by size category where the sampling programme targets only size 1 and 2 but not sizes 3-5 (more common, covered by another programme). The sizes 3-5 (in gray) are not in the sampling frame of the programme. Following cur-rent guidelines the existence of these additional sizes should not be declared to RDBES. Without any flagging of such out-of-frame strata, the estimate of size of distribution for the landing would likely be done using only data from size 1 and size 2 resulting in severe bias.

SAid	SAstratification	SAstratum	SAselectionMethod	SAnumberTotal	SAnumberSampled
1	Y	Size 1	random	2	2
2	Y	Size 2	random	6	2
---	---	Size 3	---	12	0
---	---	Size 4	---	20	0
---	---	Size 5	---	18	0

Example 4 (present situation): Simplified example of week selection from a year where 8 weeks (in gray) are outside the sampling frame of the programme (e.g., due to observer vacations). Following current guidelines the existence of these additional weeks should not be declared to RDBES. Without any flagging of this situation final estimates (e.g., total landings) could result in severe biases.

TEid	TEstratification	TEstratum	TEselectionMethod	TEnumberTotal	TEnumberSampled
1	Y	work-weeks	random	44	2
2	Y	work-weeks	random	44	2
---	---	vacation-weeks	---	8	0

Proposed solution

Having all parts of the population well identified during estimation is a necessary condition for unbiased estimation and to inform end-users on the coverage of the estimates they receive. Such documentation does not secure unbiased estimates by itself but does needs to be explicit so it can be adequately handled during estimation. As such, it is important that parts of the target population that are missing from the sampling frame are documented and quantified, both in the data *and* in the results object output from the estimation. Like the examples above document, out-of-frame parts of the population may take place at any level of the sampling hierarchy.

WKRDB-EST2 suggests that current guidelines are changed to **"when some components of the target population of each sampling level in RDBES are excluded from the sampling frame, they should to be declared and, whenever possible, also quantified".** This will enable identification of the problem, and, if declared, also the magnitude. The solution does not however give an indication on how these 'out-of-frame" should be handled in the estimation nor if the data model can adequately support an expansion from the study population to the target population. It should also be noted that there is different categories of "out-of-frame": some "out-of-frame" parts of the population will be of interest for estimates, others may not be, and ideally this distinction would be apparent in the data. This issue therefore needs discussion and testing through estimation with real sample data before a definitive conclusion can be reached.

Two possible implementations of these guidelines were considered during WKRDB-EST2: a) simple flagging by means of an additional variable in every design table (e.g., "Out of frame units present? Yes/No"), b) specification of out-of-frame components by means of an "out-of-frame" stratification row. Implementation b) was found preferred because it provides the estimator with additional evidence on the potential significance of the issue. Two variants of implementation of the latter are thought possible (displayed in red and blue in example below). Solution in red was found better for being informative and less easy to confuse with guidance for reporting missing values (selectionMethod == "not-sampled"). That implementation might also avoid calculations depending on the additional checking of column stratification to identify the component. While acknowledging the drawback of implicitly considering "out-of-frame" as a variant within selectionMethod, it is underscored that the signalling of "out-of-frame" is related to sample selection and having it identified in selectionMethod columns would significantly simplify the reading of the information on the tables.

VSid	VSstratification	VSstratum	VSselectionMethod	VSnumberTotal	VSnumberSampled
1	Y	>=10 <15m	random	500	2
2	Y	>=10 <15m	random	500	2
3	Y	>=15m	random	200	2
4	Y	>=15m	random	200	2
5	Y	<10m	Out-of-frame	1000	0
5	O	<10m	not-sampled	1000	0

If the suggested implementation is accepted, the Example 3 in documentation should be updated to

Example 3. Simplified example of stratification by vessel size where one stratum is out-of-frame (<10 m). That out-of-frame stratum is declared in RDBES with VSselectionMethod == "out-of-frame" and numberSampled set to 0. numberTotal provides an idea of the importance of out-of-frame components to the estimator.

VSid	VSstratification	VSstratum	VSselectionMethod	VSnumberTotal	VSnumberSampled
1	Y	>=10 <15m	random	500	2
2	Y	>=10 <15m	random	500	2
3	Y	>=15m	random	200	2
4	Y	>=15m	random	200	2
5	Y	<10m	out-of-frame	1000	0

Additional examples that may be worth providing in the FAQ of documentation are

Example XX. Simplified example of stratification hauls in a trip by subdivision.The sampling programme targets 27.3.a.20 but not 27.4.a. The hauls in 27.4.a (in red) are not in the sampling frame of the programme and are declared in RDBES by means of an "out-of-frame" stratum where FOselectionMethod == "out-of-frame" and numberSampled is set to 0. numberTotal provides an idea of the importance of out-of-frame components to the estimator.

FOid	FOstratification	FOstratum	FOselectionMethod	FOnumberTotal	FOnumberSampled
1	Y	27.3.a.20	random	4	2
2	Y	27.3.a.20	random	4	2
3	Y	27.4.a	out-of-frame	6	0

Example XX: Simplified example of week selection from a year where 8 weeks are outside the sampling frame of the programme (e.g., due to observer vacations). Those weeks can be declared in RDBES by means of an "out-of-frame" stratum where TEstratum == "out-of-frame" and numberSampled is set to 0 (in red). numberTotal provides an idea of the importance of out-of-frame components to the estimator.

TEid	TEstratification	TEstratum	TEselectionMethod	TEnumberTotal	TEnumberSampled
1	Y	work-weeks	random	44	2
2	Y	work-weeks	random	44	2
3	Y	vacation-weeks	out-of-frame	8	0

Example XX: Simplified example of stratification of a landing event by size category where the sampling programme targets only size 1 and 2 but not sizes 3-5 (more common, covered by another programme). The sizes 3-5 are not in the sampling frame of the programme and can be declared in RDBES by means of an "out-of-frame" stratum where SAselectionMethod == "out-of-frame" and numberSampled is set to 0 (in red). numberTotal provides an idea of the importance of out-of-frame components to the estimator.

SAid	SAstratification	SAstratum	SAselectionMethod	SAnumberTotal	SAnumberSampled
1	Y	Size 1	random	2	2
2	Y	Size 2	random	6	2
3	Y	Other sizes	out-of-frame	50	0

Issue #2 Number total and number sampled when two selection methods are used

Present situation

At the moment the RDBES documentation provides no guidelines on the possible existence of two (or more) sampling methods within a stratum. This situation may occur e.g., when countries carry out a few non-probabilistic samples in addition to their probabilistic samples with intention of compensating for data shortages such as those motivated by higher than expected non-responses.

Issue

A transcription of the gitHub issue is given below

Suppose we have a data as shown below

VSstratumName	VSencryptedVesselCode	VSselectionMethod	VSnumberTotal	VSnumberSampled
Stratum 1	Vessel 1	SRSWR	15	3
Stratum 1	Vessel 2	SRSWR	15	3
Stratum 1	Vessel 3	SRSWR	15	3
Stratum 1	Vessel 4	NPAH	15	2
Stratum 1	Vessel 5	NPAH	15	2

In the example there is one stratum having two selection methods. One method is probabilistic, another is ad-hoc. What is the correct way of registering the number of vessels sampled? Should this number be calculated for every selection method separately, 3 for SRSWR and 2 for NPAH? Or maybe 5 should be entered in all rows as the total number of vessels sampled regardless of the selection method.

Adding to this, subGroup 1 of WKRDB-EST2 also reported:

Now it's possible to submit the data with several different selection methods within each stratum. Is it statistically ok? What if there is e.g. SRSWR and NPAH method in one stratum. Should the numberTotal and numberSampled include only the units probabilistically sampled or all the units that were sampled in this stratum (as is said in the description of the column in the data model). How are we going to carry out all the estimation in this case?

Analysis

The existence of two sampling methods and sampling intensities in a single stratum is frequent when e.g., substitution sampling is used to attenuate non-responses. However, its declaration in the RDBES is at present confusing and may lead to errors in the calculations of sampling probabilities.

The proposal of solutions for that situation requires the consideration of a) how non-probabilistic methods should be handled when reporting and estimating using RDBES and b) adjustments that may be needed to sampling probabilities in its presence. Topic a) is presently being considered by a subgroup of WGCATCH which is discussing the reporting of non-probabilistic events in the RDBES. Topic b) is yet to be fully considered in the context of the RDBES estimation.

Expanding on the example given, the following case-studies regarding this issue were considered at WKRDB-EST2:

- Two NPAH events added as a supplement to the original units sampled via SRSWOR. Only SRSWOR samples are considered "representative" of the population.
- Two NAPH events added as a supplement to the original units sampled via SRSWOR. All SRSWOR samples are considered "representative" of the population.

- Two of the SRSWOR units were non-responses that were later replaced with two NPAH events considered "representative" of the population.
- Two of the SRSWOR units were non-responses that were later replaced with two NPAH events that, however, are considered "non-representative" of the population.

In the alternatives above, considerations such as "representative" and "non-representative" are really about communication between the "data submitter" and the "data estimator": they are a means for the data submitter to highlight to the data estimator one's opinion on the future handling of those non-probabilistic samples during estimation. Such opinion is not binding (in the context of RDBES the "estimator" can always overrule it and leave the options made documented) but will still be useful from the point of view of establishing default settings for many of the functions developed.

In the analysis of the examples below, WKRDB-EST2 participants highlight that the discussion of the handling of non-responses and adjustments of sampling probabilities are at its infancy within the RDBES development community and the ICES commercial catch sampling community in general. To the knowledge of WKRDB-EST2 chairs, non-responses are not routinely handled by ices countries when carrying out their annual estimates. Examples here provided should therefore be taken with a "pinch of salt", mostly as documentation of a kick-off discussion which details will for sure be reanalyzed by the core-group of development of the RDBES and by future editions of WGCATCH and WKRDB-EST.

Two NPAH events added as a supplement to the original units sampled via SRSWOR. Only SRSWOR samples are considered "representative" of the population.

Table below displays a simplified view of the RDBES under case-study circumstances. Column "Representative" is not presently in RDBES data model and is added here just as a clarification of the interpretation suggested by the data submitter to the estimator with regards to the use of those data. "Design weight original" and "Design weight adjusted for NR" represent the calculations of design weights implicit to the suggestion of the data submitter before and after consideration of non-responses. In this case there were no non-responses so design weights with and without their consideration are the same and can be obtained from the count of events sampled AND representative. The same value can *only be derived* from values in VSnumberSampled if these are set to "3".

VSid	VSstratum	VSencryptedVesselCode	VSselectionMethod	VSnumberTotal	VSnumberSampled	VSsampled	*Representative*	*Design weight original*	*Design Weight adjusted for NR*
1	Stratum 1	vsl_1	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/3*
2	Stratum 1	vsl_2	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/3*
3	Stratum 1	vsl_3	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/3*
4	Stratum 1	vsl_4	NPAH	15	2? 5?	Y	*N*	*0*	*0*
5	Stratum 1	vsl_5	NPAH	15	2? 5?	Y	*N*	*0*	*0*

Two NAPH events added as a supplement to the original units sampled via SRSWOR. All SRSWOR samples are considered "representative" of the population.

Table below displays a simplified view of the RDBES under case-study circumstances. Column "Representative" is not presently in RDBES data model and is added here just as a clarification of the interpretation suggested by the data-submitter to the estimator with regards to the use of those data. "Design weight original" and "Design weight adjusted for NR" represent the calculations of design weights implicit to the suggestion of the data submitter before and after consideration of non-responses. In this case there were no non-responses so design weights with and without their consideration can be obtained from the count of events sampled AND representative. The same value can only be generated from the values in VSnumberSampled if these are set to "5".

VSid	VSstratum	VSencryptedVesselCode	VSselectionMethod	VSnumberTotal	VSnumberSampled	VSsampled	*Representa-tive*	*Design weight original*	*Design Weight adjusted for NR*
1	Stratum 1	vsl_1	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/5*
2	Stratum 1	vsl_2	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/5*
3	Stratum 1	vsl_3	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/5*
4	Stratum 1	vsl_4	NPAH	15	2? 5?	Y	*Y*	*0*	*15/5*
5	Stratum 1	vsl_5	NPAH	15	2? 5?	Y	*Y*	*0*	*15/5*

Two of the SRSWOR units were non-responses that were later replaced with two NPAH events considered "representative" of the population.

Table below displays a simplified view of the RDBES under case-study circumstances. Column "Representative" is not presently in RDBDES data model and is added here just as a clarification of the interpretation suggested by the data-submitter to the estimator with regards to the use of those data. "Design weight original" and "Design weight adjusted for NR" represent the calculations of design weights implicit to the suggestion of the data submitter before and after consideration of non-responses. Table demonstrates that value "3" present in the denominator of design weights adjusted for non-response that are required for simple estimation of stratum totals can be derived from the count of events both sampled AND representative. It can also be generated from the values in VS sampled if these are set to "3".

VSid	VSstratum	VSencryptedVesselCode	VSselectionMethod	VSnumberTotal	VSnumberSampled	VSsampled	*Representative*	*Design weight original*	*Design Weight adjusted for NR*
1	Stratum 1	vsl_1	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/3*
2	Stratum 1	vsl_2	SRSWOR	15	3? 5?	N	*Y*	*15/3*	*0*
3	Stratum 1	vsl_3	SRSWOR	15	3? 5?	N	*Y*	*15/3*	*0*
4	Stratum 1	vsl_4	NPAH	15	2? 5?	Y	*Y*	*0*	*15/3*
5	Stratum 1	vsl_5	NPAH	15	2? 5?	Y	*Y*	*0*	*15/3*

Two of the SRSWOR units were non-responses that were later replaced with two NPAH events that however are considered "non-representative" of the population.

Table below displays a simplified view of the RDBES under case-study circumstances. Column "Representative" is not presently in RDBDES data model and is added here just as a clarification of the interpretation suggested by the data-submitter to the estimator with regards to the use of those data. "Design weight original" and "Design weight adjusted for NR" represent the calculations of design weights implicit to the suggestion of the data submitter before and after consideration of non-responses. Table demonstrates that value "1" present in the denominator of design weights adjusted for non-response that are required for simple estimation of stratum totals can be derived from the count of events both sampled AND representative *but not* from the values in VS sampled.

VSid	VSstratum	VSencryptedVesselCode	VSselectionMethod	VSnumberTotal	VSnumberSampled	VSsampled	*Representa-tive*	*Design weight original*	*Design Weight adjusted for NR*
1	Stratum 1	vsl_1	SRSWOR	15	3? 5?	Y	*Y*	*15/3*	*15/1*
2	Stratum 1	vsl_2	SRSWOR	15	3? 5?	N	*Y*	*15/3*	*0*
3	Stratum 1	vsl_3	SRSWOR	15	3? 5?	N	*Y*	*15/3*	*0*
4	Stratum 1	vsl_4	NPAH	15	2? 5?	Y	*N*	*0*	*0*
5	Stratum 1	vsl_5	NPAH	15	2? 5?	Y	*N*	*0*	*0*

Suggestions to core-group

- The case-study analysis indicated that VSnumberSampled may not be useful for the calculation of probability of inclusions and design weights under SRSWOR, when non-responses are present. Rather a count of VSsampled==Y is needed. The exact number displayed in VSnumberSampled seems therefore to be more of interest to other uses of the data (e.g., reporting number of samples, etc) than for design-based estimation where it seems to only be applicable when non-responses are absent and data is collected probabilistically. As it is confusing to have 2 different values reported for sample size of one single stratum, the subgroup considers that it would be better to define "numberSampled" as the total number of attempts made at sampling. In the present case, this would mean "5" would be written in that column in any of the case-studies above represented.
- Before any final conclusion is made on this issue, a look will be needed into how the selection probability and inclusion probability should be declared in these cases and also for cases with more complicated selection methods e.g. UPSWOR. The probabilities are the ones being relevant for estimation and therefore also the place where handling of a specific case can be communicated. It may very well be that the conclusion is that it is beneficial to include in the data model "adjusted probabilities" that taking the non-response issue into account along other adjustments of the probabilities.
- By not including any indication on "representativeness" of non-probabilistic events, v1.18 of RDBES has difficulties in providing estimators with the info needed to handle these events during the estimation. In practice, this equates to an implicit option for having estimators consider, as a default, that all non-probabilistic data in RDBES is non-representative and should therefore be excluded in analysis. Albeit statistically reasonable, the pros and cons of such default option should be carefully considered. It seems to the WKRDB-EST2 subgroup analysing this issue that the latter option would be more cumbersome and that data-estimators might not always be in position of making that decision, particularly when handling data from other countries or raising data from regional designs. It is therefore advisable to include some sort of representativeness indicator/opinion alongside the data.
- The model displayed above for conveying representativeness is only one among other alternatives that can be considered. Alternatives like including information on representativeness at the level of selection-method codes are also possible. Evaluations of the relative merit of different alternatives should consider the easiness of their application to the calculation of adjusted weights.
- The "representative" info gives autonomy and responsibility to data submitter allowing him/her to guide the default estimation, but the estimator can always decide otherwise and take the responsibility of including "non-representative" events in the estimation by oneself.
- Case-study analysis and weight adjustments presented above assumes non-responses are missing completely at random. That is a simple and useful default approach to be used as starting point for developing of RDBES functions. But it is likely not the situation in the majority of cases. A vast array of other alternatives are possible with the RDBES data model that involve different types of auxiliary data and can improve accuracy of final estimates.

Issue #3 Joint inclusion probabilities from unequal probability sampling

Issue

In simpler cases, e.g. SRSWOR or SRSWR, all inclusion probabilities needed for variance estimation can be calculated from sample size and population size. However, joint inclusion probabilities are required for estimation of variance for unequal probability designs such as UPSWOR. These are not currently incorporated into the RDBES format and would require either repetition of rows or the inclusion in the data model of the possibility of adding matrices of joint inclusion probabilities for units within a sample.

Related issue: https://github.com/ices-tools-dev/RDBES/issues/76

Proposed solution

Estimating unequal probability designs requires advanced knowledge of survey statistics. WKRDB-EST2 proposes that joint inclusion probabilities are not incorporated into the RDBES and that institutes using these more complicated survey types are allowed to supply these types of additional information required for estimation as separate formats. This guidance should be added to the RDBES documentation in its next update.

Issue #4 Usage of SSuseCalcZero

Present situation

SSuseCalZero variable present in the data model is presently defined as *Indicating if the data can be used for calculating zeros 'Yes' or 'No'. E.g. used by Denmark for sampling targeting specific stocks.* The rationale behind it is that, when SSuseCalZero == "N", zeros will not be generated for the species on the species list that do not appear in the SA table. And that when SSuseCalZero == "Y", zero generation can proceed.

Issue

Present documentation is scarce in details on how to handle situations like incomplete "concurrent sampling" where not all species present in a landing event or fishing operation were quantified and zeros should not be calculated for species present in the species but missing from the sampling data.

Suggested solution

The core-group of RDBES development is suggested to add the following text and FAQs below to the documentation

Under concurrent sampling, SSuseCalcZero should be used like a quality indicator - If data submitter reports "Yes" then concurrent sampling will be assumed finished and zeros will be calculated for the species missing in the SA table. If data submitter reports "No" then concurrent sampling will be assumed not finished and zeros will not be added to the species in the species list that are absent from SA records. In the latter situation SSnumTotal should be reported as NA.

FAQ: During concurrent sampling of landings or discards, I did not finish the sampling of all species in my species list due to e.g., time limitations. How should I report this situation?

R: Set SSuseCalcZero to "No" so that zeros are not calculated for the remainder species you did not collect information on.

FAQ: If I select the species I will sample (e.g., via SRSWOR), how should I report the variable SSuseCalcZero?

R: Set SSuseCalcZero to "No" to avoid that the species you did not select are not assumed absent during calculations.

Issue #5 Declaration of species*size combinations in SA table

Issue

The following example was provided for this issue:

In an onboard trip, we sample landings and discards. Landings are sampled by size category. Discards sampled by taking 3 baskets. Baskets are pooled together before sorting.

The way I interpret the design is that the haul is stratified into landings and discards. Then:

- *landings are stratified into species. And species further stratified into size categories. Samples are then taken from each size category.*
- *discards are sampled (baskets). Then the content of the baskets is stratified into species. And a sample taken from each species.*

I see two possibilities to declare this design with regards to the landings component and wonder which one is better

- *Option 1: declare a landings row only in the SS table*
  -- SA table starts directly with species landed with stratification ==Y
  --- size categories are children of species rows, with stratification == Y

- *Option 2: declare a landings row in SS table and in the SA table*
  -- SA table starts with a landings row with stratification == N
  --- species are children of that landings row, with stratification==Y
  --- size categories are children of species rows, with stratification == Y
- *Option 3: declare a landings row in the SS table, and then just have an SA row for each combination of species and size category, with no sub-samples. E.g. if one sampled unsorted cod, and 3 size categories of haddock then one would just have 4 SA rows (COD unsorted, HAD size 1, HAD size 2, HAD size 3).*

The user provides the following opinion

Seems to me that option 2 is more explicit and also allows the entering of total weight of landings while option 1 does not. But it is also more complex involving similar work to the discard case.

The user further suggests

a FAQ might be needed if one of the alternatives is found to allow for correct estimation. Specific code for estimating both alternatives might need to be considered if both are considered to work and choice is left to the user

Analysis

The following examples were produced at WKRDB-EST2

Alternative 1

SSid	SScatchFra
1	Lan
2	Dis

stratification by spp, then some spp by size

Ssid	SArecType	SAseqNum	SAparSequNum	SAstratification	SAstratumName	SAspeCode	SAcatchCat	SAlandCat	SAunitType	SAtotalWtLive	SAsampWtLive	SAnumTotal	SAnumSamp	SAselectMeth	SAunitName	SAlowHierarchy	SAsamp
1	SA	1	NA	Y	cod	cod	Lan		box	320	60	8	3	SRSWOR	box 1	A	Y
1	SA	2	NA	Y	Had	had	Lan		box	2700	2700	145	145	CENSUS	pop had	NA	Y
1	SA	3	2	Y	size 1	had	Lan		box	100	100	5	5	SRSWOR	box 2	A	Y
1	SA	4	2	Y	size 2	had	Lan		box	600	100	30	5	SRSWOR	box 3	A	Y
1	SA	5	2	Y	size 3	had	Lan		box	1000	100	50	5	SRSWOR	box 4	A	Y
1	SA	5.1	2	Y	other	had	Lan		box	1000	0	60	0	not-sampled		D	N
2	SA	6	NA	Y	all discards	Animalia	Dis		baskets			30	1	SRSWOR	basket 1	D	Y
2	SA	7	6	Y	cod	cod	Dis		portion			0,5	0,5	SRSWOR	basket 1	D	Y
2	SA	8	6	Y	had	had	Dis		portion			0,5	0,5	SRSWOR	basket 2	D	Y

Alternative 2

SSid	SScatchFra
1	Lan
2	Dis

Ssid	SArecType	SAseqNum	SAparSequNum	SAstratification	SAstratumName	SAspeCode	SAcatchCat	SAlandCat	SAunitType	SAtotalWtLive	SAsampWtLive	SAnumTotal	SAnumSamp	SAselectMeth	SAunitName	SAlowHierarchy	SAsamp
1	SA	1	NA	N	all landings	Animalia	Lan		box	3020	3020	153	153	CENSUS	land 1	NA	Y
1	SA	2	1	Y	cod	cod	Lan		box	320	60	8	3	SRSWOR	box 1	A	Y
1	SA	3	1	Y	Had	had	Lan		box	2700	2700	145	145	CENSUS	pop had	NA	Y
1	SA	4	3	Y	size 1	had	Lan		box	100	100	5	5	SRSWOR	box 2	A	Y
1	SA	5	3	Y	size 2	had	Lan		box	600	100	30	5	SRSWOR	box 3	A	Y
1	SA	6	3	Y	size 3	had	Lan		box	1000	100	50	5	SRSWOR	box 4	A	Y
1	SA	5.1	3	Y	other	had	Lan		box	1000	0	60	0	not-sampled		D	N
2	SA	7	NA	N	all discards	Animalia	Dis		baskets			30	1	SRSWOR	basket 1	D	Y
2	SA	8	7	Y	cod	cod	Dis		portion			0,5	0,5	SRSWOR	basket 1	D	Y
2	SA	9	7	Y	had	had	Dis		portion			0,5	0,5	SRSWOR	basket 2	D	Y

Alternative 3

SSid	SScatchFra
1	Lan
2	Dis

Ssid	SArecType	SAseqNum	SAparSequNum	SAstratification	SAstratumName	SAspeCode	SAcatchCat	SAlandCat	SAunitType	SAtotalWtLive	SAsampWtLive	SAnumTotal	SAnumSamp	SAselectMeth	SAunitName	SAlowHierarchy	SAsamp
1	SA	1	NA	Y	cod unsorted	cod	Lan		box	320	60	8	3	SRSWOR	box 1	A	Y
1	SA	2	NA	Y	Had size 1	had	Lan		box	100	100	5	5	SRSWOR	box 2	A	Y
1	SA	3	NA	Y	Had size 2	had	Lan		box	600	100	30	5	SRSWOR	box 3	A	Y
1	SA	4	NA	Y	Had size 3	had	Lan		box	1000	100	50	5	SRSWOR	box 4	A	Y
1	SA	4.1	NA	Y	Other	had	Lan		box	1000	0	60	0	not-sampled		D	N
2	SA	5	NA	Y	all discards	Animalia	Dis		baskets			30	1	SRSWOR	basket 1	D	Y
2	SA	6	5	Y	cod	cod	Dis		portion			0,5	0,5	SRSWOR	basket 1	D	Y
2	SA	7	5	Y	had	had	Dis		portion			0,5	0,5	SRSWOR	basket 2	D	Y

Alternative 4

SSid	SScatchFra
1	Lan
2	Dis

Ssid	SArecType	SAseqNum	SAparSequNum	SAstratification	SAstratumName	SAspeCode	SAcatchCat	SAlandCat	SAunitType	SAtotalWtLive	SAsampWtLive	SAnumTotal	SAnumSamp	SAselectMeth	SAunitName	SAlowHierarchy	SAsamp
1	SA	1	NA	Y	cod	cod	Lan		box	320	60	8	3	SRSWOR	box 1	A	Y
1	SA	3	2	Y	size 1	had	Lan		box	100	100	5	5	SRSWOR	box 2	A	Y
1	SA	4	2	Y	size 2	had	Lan		box	600	100	30	5	SRSWOR	box 3	A	Y
1	SA	5	2	Y	size 3	had	Lan		box	1000	100	50	5	SRSWOR	box 4	A	Y
1	SA	5.1	2	Y	size 4	had	Lan		box	1000	0	60	0	not-sampled		D	N
2	SA	6	NA	Y	cod	cod	Dis		portion			0,5	0,5	SRSWOR	basket 1	D	Y
2	SA	7	NA	Y	had	had	Dis		portion			0,5	0,5	SRSWOR	basket 2	D	Y

Suggested solution

WKRDB-EST2 recognized that all the above alternatives are valid and that flexibility is positive for data uploads given the present differences in sub-sampling and the data models of databases used at national level. Still from a data usage perspective, alternative 1 was concluded to be the one that should be suggested to users. The grounds for this conclusion were:

- Alternative 1 and Alternative 2 differ in the presence of a "top Animalia" row associated with the "landings". WKRDB-EST2 concluded that row is only needed when there is sampling of the content of a fraction (as it happens in the case of discards); it is not needed when the sampling methodology is CENSUS (as happens in the case of landings). Alternative 1 is therefore preferable in relation to alternative 2.
- Alternative 1 and alternative 3 differ in the presence in alternative 1 of a top row summarizing the totals for species HAD. WKRDB-EST2 concluded that top row to be useful - it provides information on what is missing by letting the user know that it is a stratum within HAD that is missing and not, e.g., a species not sampled. This will be impossible under alternative 3 where the missing values reported will impact the entire landings. Alternative 1 is therefore preferable to alternative option 3.
- Alternative 1 and alternative 4 differ in the declaration of the "top Animalia" row for discards in alternative 1 (absent in alternative 4). Under the presence of subsampling taking place on the bulk of discards (as in the example given), alternative 1 is to be preferred; Alternative 4 does not leave clear that there was a subsampling of the bulk and involves considerable assumptions in the calculation of sampling probabilities that are not explicitly stated - it should only be used when a census of discards take place (see comparison of alternative 1 and alternative 2).

Issue #6 Using minutes as sampling units, how to calculate sampling probabilities and % of time covered

Issue

https://github.com/ices-tools-dev/RDBES/issues/74

Analysis

Due to time constraints it was not possible to conclude on this issue during WKRDB-EST2.

Suggestion

Participants in WKRDB-EST2 point out that Annex 10 of WKRDB-EST report appears to have a somewhat analogue situation. That annex should be the starting point to further consideration of this issue.

Issue #7–9

Due to time constraints these issues were not discussed during WKRDB-EST2.