

RDBqc R package and Automatic reporting tools

Second Training on quality checks
15/04/2024

Walter Zupa



COISPA



COISPA



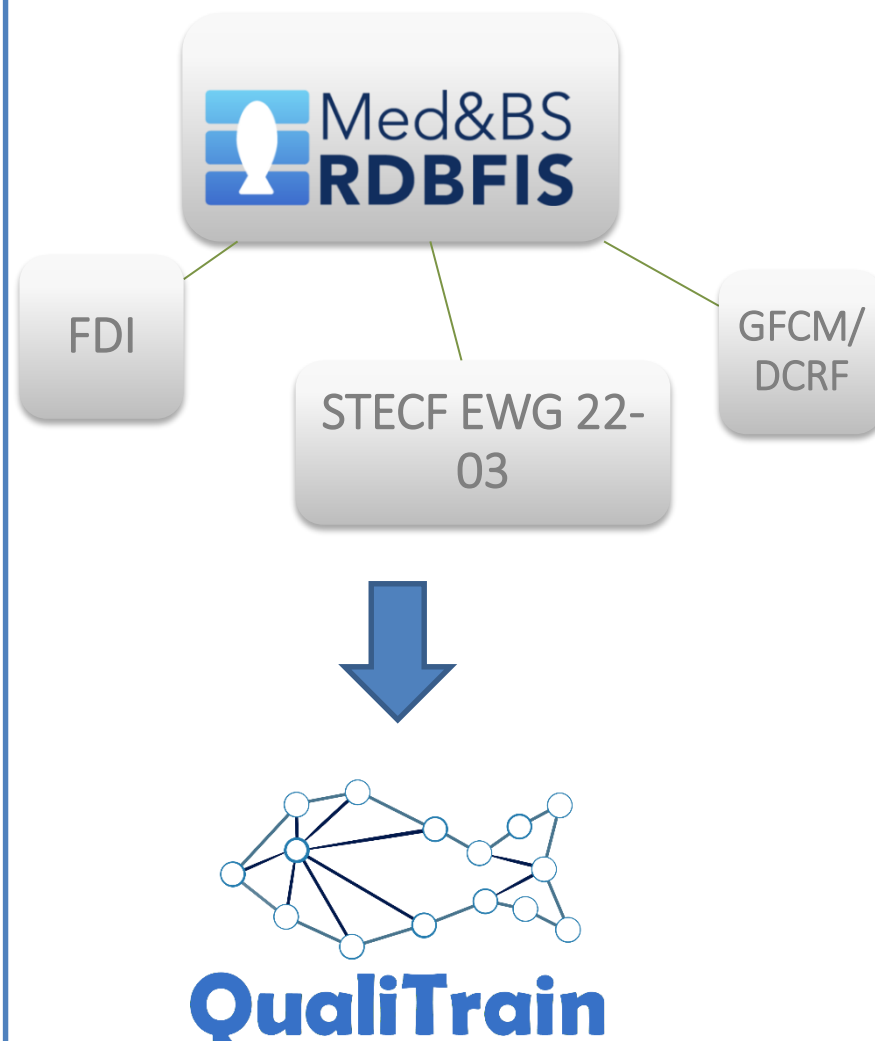
CNR
IRBIM
Istituto per lo studio e la
ricerca in biologia marina
e acquacoltura



QualiTrain takes advantage from existing quality check tools developed and used in different contexts.

In particular, the RDBqc R package developed in **MED&BS RDBFIS** project for consistency quality checks, including specific quality checks for all the relevant formats utilized for data exchanging in Mediterranean and Black Sea (detailed and aggregated) was specifically adopted.

RDBqc package has been **integrated in the RDBFIS** but is also available as a **stand-alone tool** to be used by MS before the upload of the data on the RDBFIS for a timely detection of errors.



Data validation and quality checking tools implemented in RDBqc

Supported data call formats:

- Med&BS datacall (<https://datacollection.jrc.ec.europa.eu/dc/medbs>);
- FDI datacall (<https://datacollection.jrc.ec.europa.eu/dc/fdi>);
- GFCM DCRF data call (<http://www.fao.org/gfcm/data/dcrf/platform/en/>);
- RCG (https://datacollection.jrc.ec.europa.eu/bg_BG/docs/rcg)
- Cross-checks among data calls (developed in QualiTrain project)

For each table and data call:

1. Syntactic and conformity checks;
2. Data consistency quality checks.

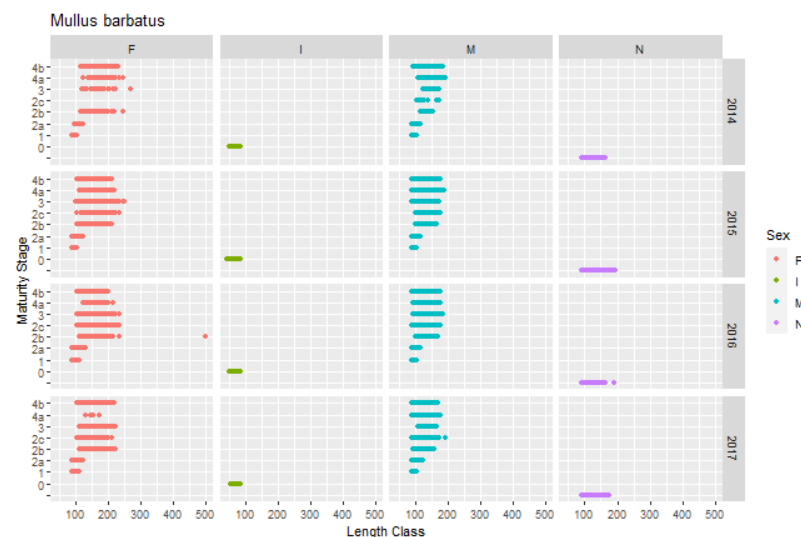
Structure of the RDBqc package

- ~ 109 functions included in the package
- Dummy datasets included to run executable examples
- Readme file and vignettes to support users in the use of functions included in the package
- Functions' documentation is embedded in the package
- Updated version of the software is publically available can be downloaded from a GitHub repository

check maturity stages

This function `RCG_check_mat` plots for each year and sex the length-maturity stage scatter plot

```
RCG_check_mat(data_ex, MS="ITA", GSA="GSA99", SP="Mullus barbatus")
```



GitHub

<https://github.com/COISPA/RDBqc>

Environment

R version > 4.1.0 (4.2.2)

Rstudio: v. 23.12.0 – 369

License: GPL-3

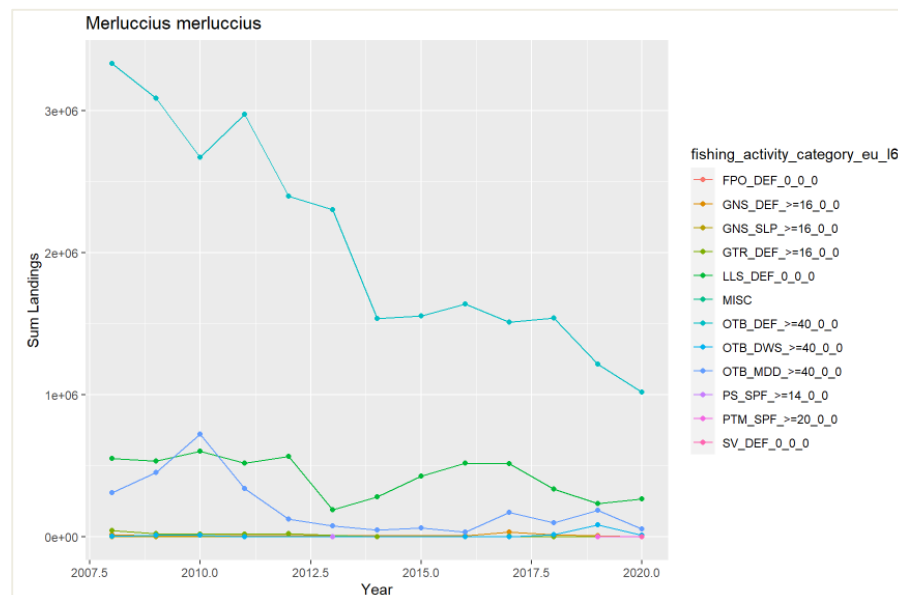


RDBqc: data quality checks for RCG format (CS and CL)

Table	Check function
CL	check_CL
CS	check_AL
CS	check_LFD
CS	check_LFD_comm_cat
CS	check_lw
CS	check_mat
CS	check_mat_ogive
CS	summarize_trips
CS	summarize_ind_meas
CS	check_loc

Checks on CL table:

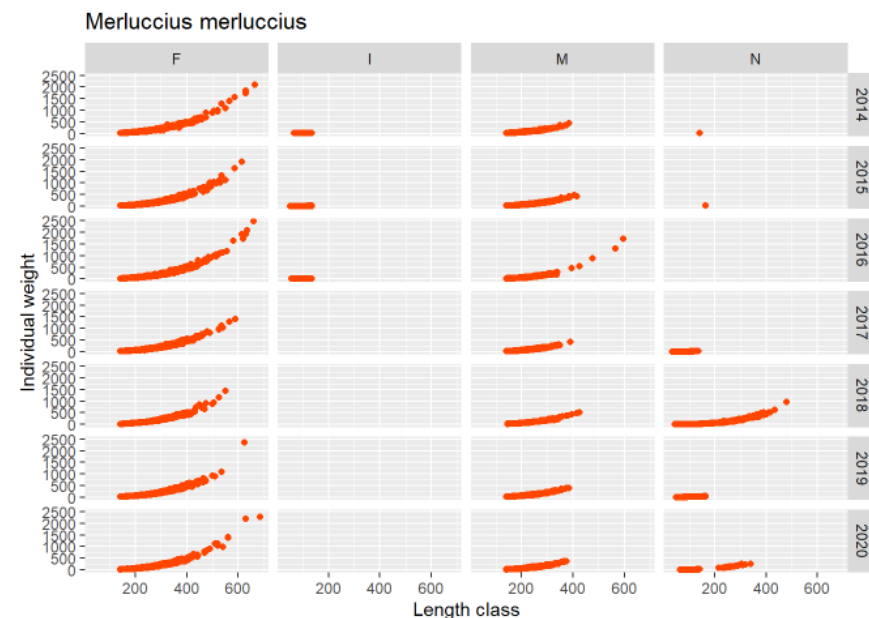
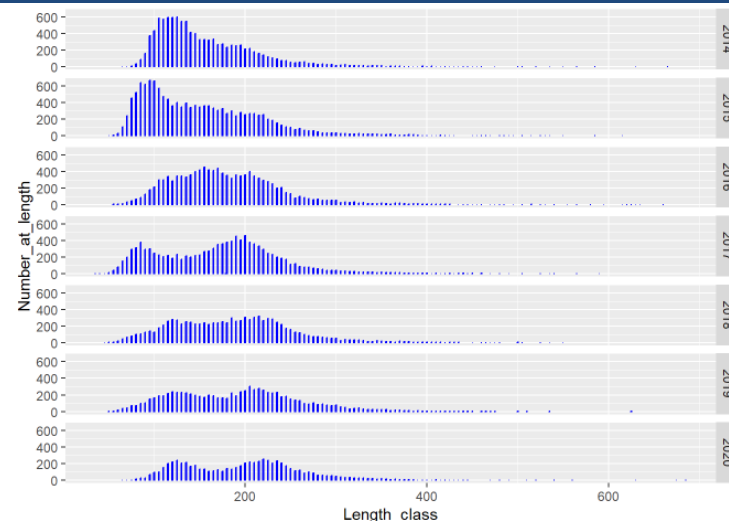
- temporal coverage of landing and landing value data (according to different timeframes: month, quarter, year)
- spatial coverage (GSA and harbour) by year
- coverage of species and métier, qualitative checks of time



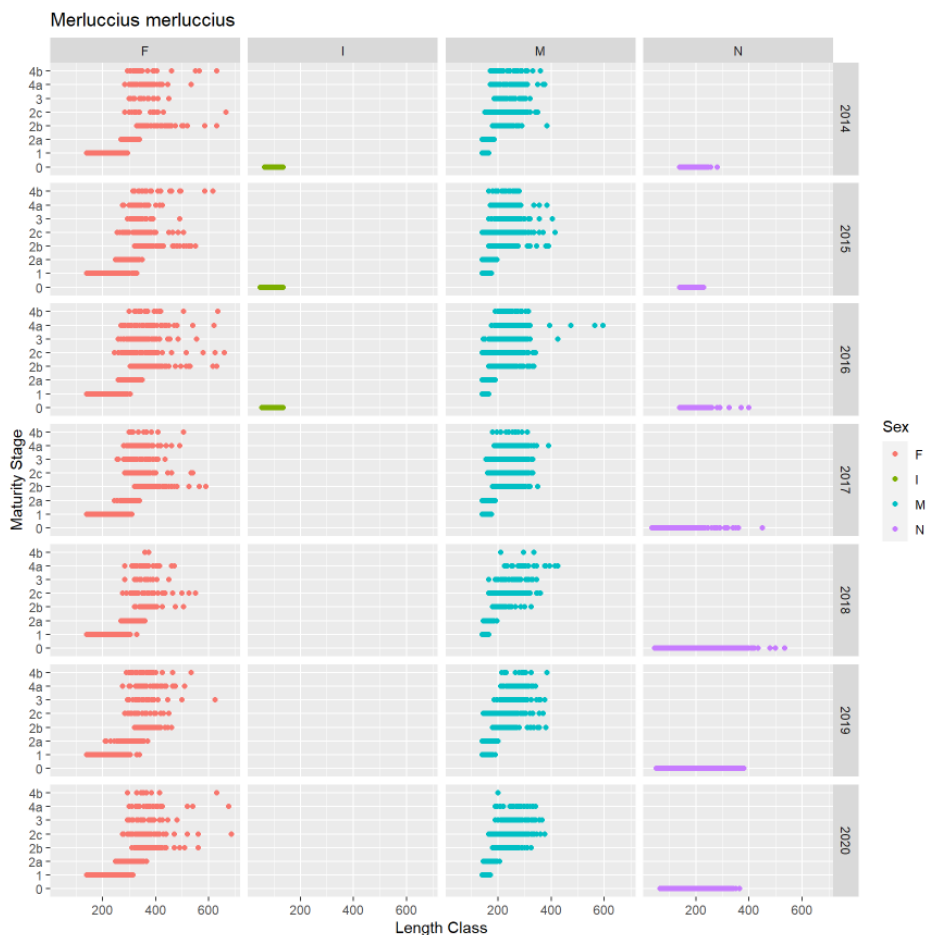
RDBqc: data quality checks for RCG format (CS and CL)

Consistency (on CS) of:

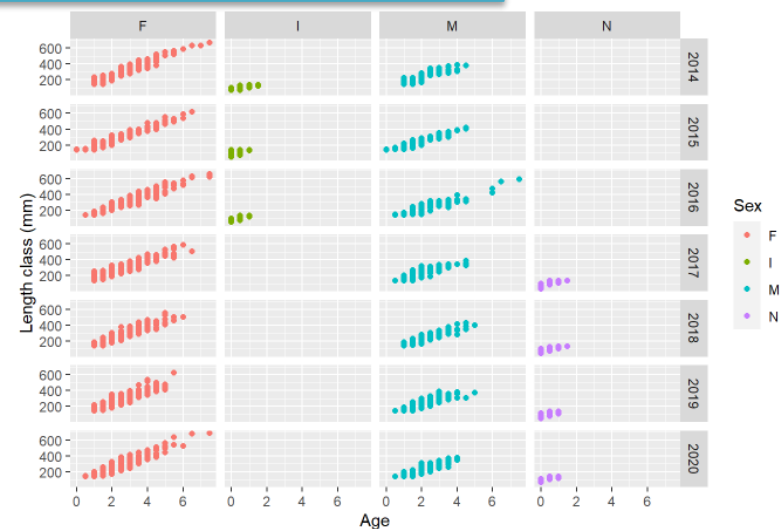
- LFD by year according to commercial category;
- length-weight data;
- age-length data;
- sex-length-maturity stage;
- Summary of the number of individual biological data (length, sex, maturity, weight and age) collected by trip;
- Summary of the number of trips/hauls monitored by year, port, metier, sampling method;



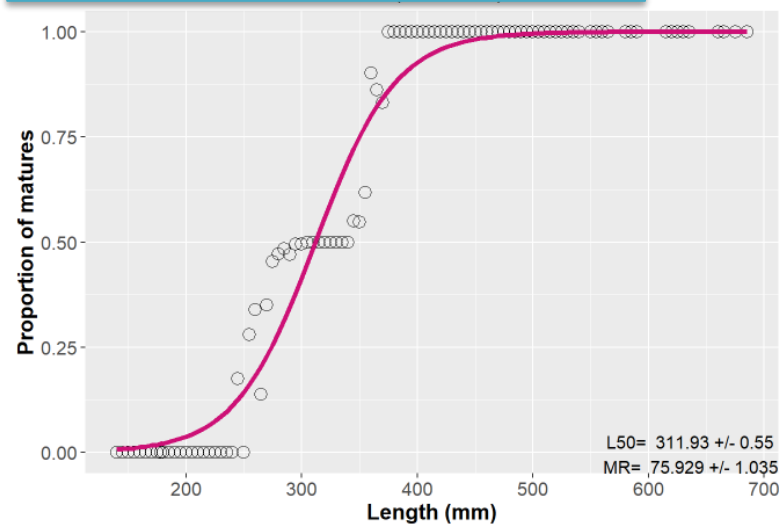
Consistency of sex and maturity stage



Check of age-length data



Check of maturity ogive by sex



RDBqc: data quality checks for RCG format (CS and CL)

2 Detailed data table

2.1 Summary of trips

Biological data were analysed to provide a summary table reporting the number of trips/hauls monitored by year, port, metier and sampling method.

2.1.1 GSA18

2.1.1.1 Merluccius merluccius

Summary table reporting the number of trips/hauls monitored by year by port, metier, sampling method for Merluccius merluccius in GSA18

Year	Flag_country	Area	Harbour	Species	Fishing_activity_category_European_IV_6	Sampling_method	Nb_trips
2014	ITA	GSA18	ITBCE	Merluccius merluccius	OTB_DEF_>=40_0_0	Observer	2
2014	ITA	GSA18	ITBCE	Merluccius merluccius	OTB_DEF_>=40_0_0	SelfSampling	2
2014	ITA	GSA18	ITBDS	Merluccius merluccius	OTB_DEF_>=40_0_0	Observer	4

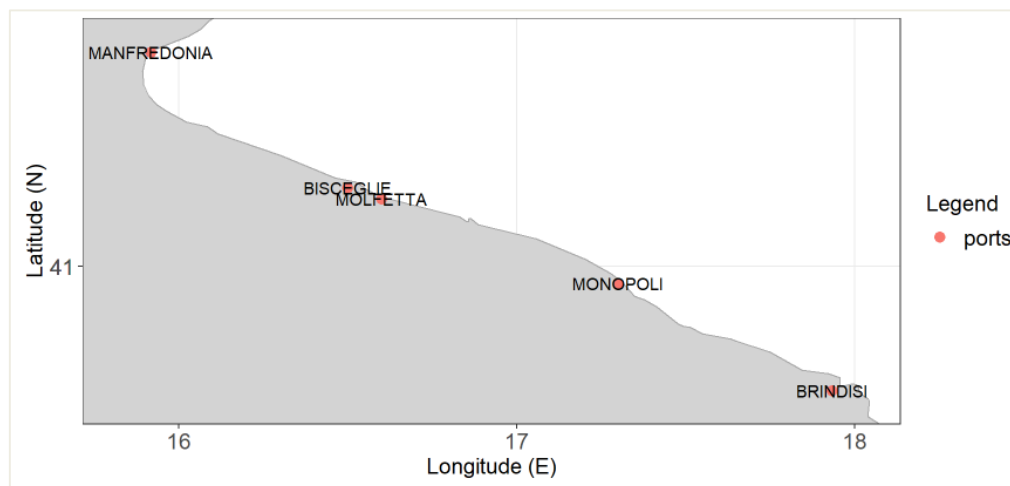
CS table

SUMMARY TABLE

- Number of trips monitored by year, port, metier, sampling method

PLOT

- Map of data spatial distribution



RDBqc: data quality checks for Med&BS data call

Table	Check function	Table	Check function
Catch	MEDBS_check_duplicates	Discard	MEDBS_check_missing_years
Catch	MEDBS_Catch_coverage	Discard	MEDBS_comp_disc_YQ
Catch	MEDBS_SOP	Discard	MEDBS_comp_disc_YQ_fishery
Catch	MEDBS_check_disaggregated	Discard	MEDBS_disc_mean_weight
Catch	MEDBS_check_missing_years	Discard	MEDBS_LFD
Landing	MEDBS_check_duplicates	Discard	MEDBS_ks
Landing	MEDBS_check_disaggregated	Discard	MEDBS_length_ind
Landing	MEDBS_check_missing_years	Discard	MEDBS_lengthclass_0
Landing	MEDBS_Landing_coverage	Discard	MEDBS_plot_disc_vol
Landing	MEDBS_comp_land_Q_VL_fishery	Discard	MEDBS_plot_discard_ts
Landing	MEDBS_comp_land_YQ_fishery	Discard	MEDBS_weight_0
Landing	MEDBS_ks	Discard	MEDBS_weight_minus1
Landing	MEDBS_lengthclass_0	Discard	MEDBS_yr_missing_length
Landing	MEDBS_length_ind	GP	MEDBS_check_duplicates
Landing	MEDBS_LFD	GP	MEDBS_GP_check
Landing	MEDBS_yr_missing_length	GP	MEDBS_LW_check
Landing	MEDBS_weight_0	MA	MEDBS_check_duplicates
Landing	MEDBS_weight_minus1	MA	MEDBS_MA_check
Landing	MEDBS_land_mean_weight	ML	MEDBS_check_duplicates
Landing	MEDBS_plot_land_vol	ML	MEDBS_ML_check
Landing	MEDBS_plot_landing_ts	SRA	MEDBS_check_duplicates
Landing	MEDBS_comp_land_Q_VL	SRA	MEDBS_SA_check
Landing	MEDBS_comp_land_YQ	SRL	MEDBS_check_duplicates
Discard	MEDBS_Discard_coverage	SRL	MEDBS_SL_check
Discard	MEDBS_check_duplicates	ALK	MEDBS_check_duplicates
Discard	MEDBS_check_disaggregated	ALK	MEDBS_ALK

STREAM and STECF EWG 21-02 and 22-03 included

The functions work using:

- common filters: GSA, MS, SP
- verbose parameter

Checks of duplicates: covering all the tables

Checks of disaggregated data: checks the presence of records with the same stratum but with different volumes of landings and discards (in landings, discards and catch tables)

Checks on coverage: plot and table of the volumes time series in landings, discards and catch tables

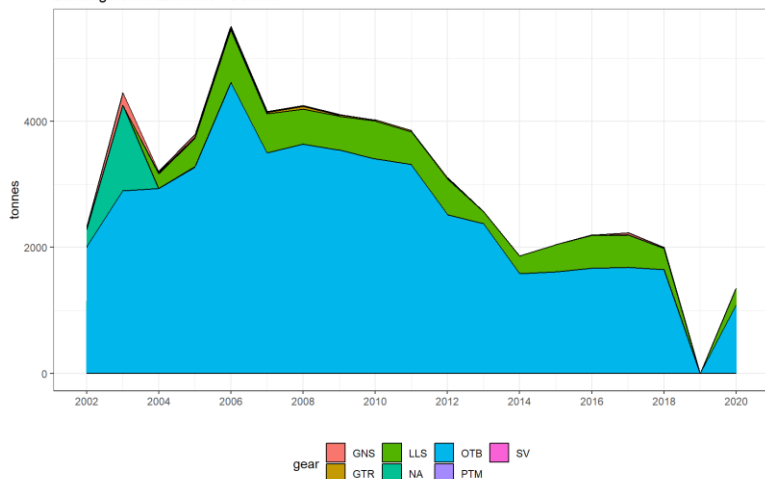
Consistency of time series: check the presence of missing years (in landings, discards and catch tables)

Weights comparison: Compare weights at different level of aggregations

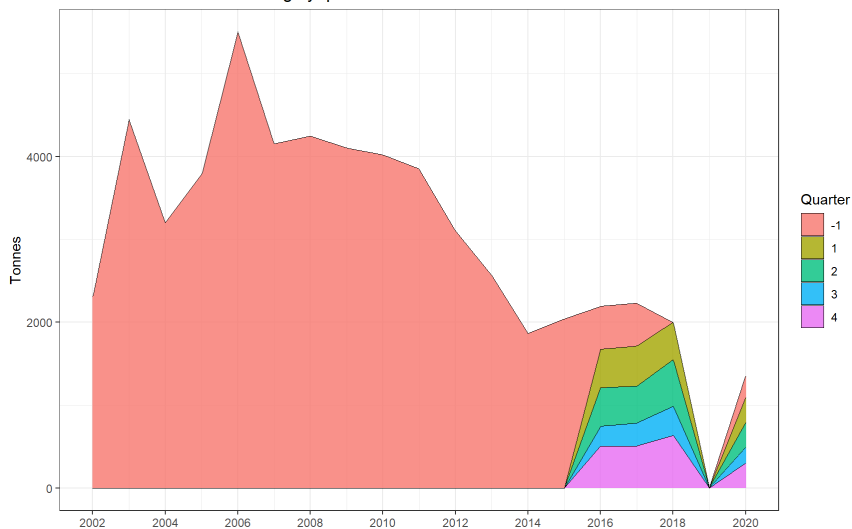
- By gear and quarters: data with VL information vs no VL information (in landings)
- By fishery and quarters: data with VL information vs no VL information (in landings)
- By gear : data by quarter vs data by year (in landings and discards)
- By gear and fishery: data by quarter vs data by year (in landings and discards)

Total landings by year and by quarters

Landings of HKE in ITA - GSA 18

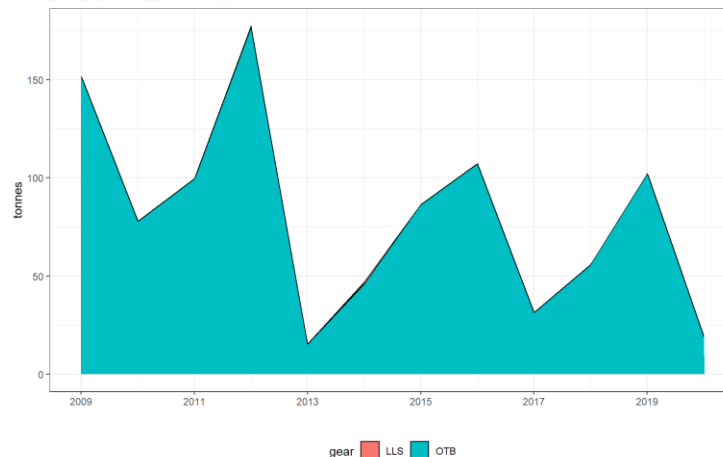


HKE ITA GSA 18 Total landing by quarter

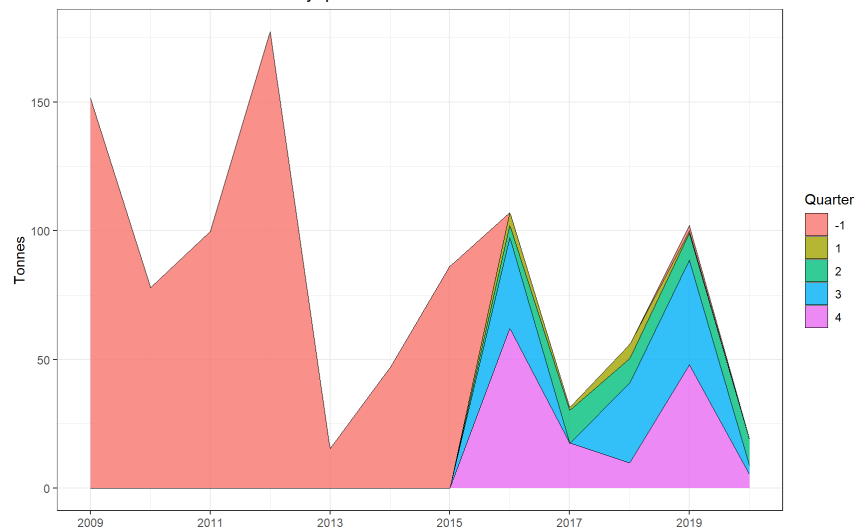


Total discards by year and by quarters

Discards of HKE in ITA - GSA 18



HKE ITA GSA 18 Total discard by quarter



The functions work using:

- common filters: GSA, MS, SP
- verbose parameter

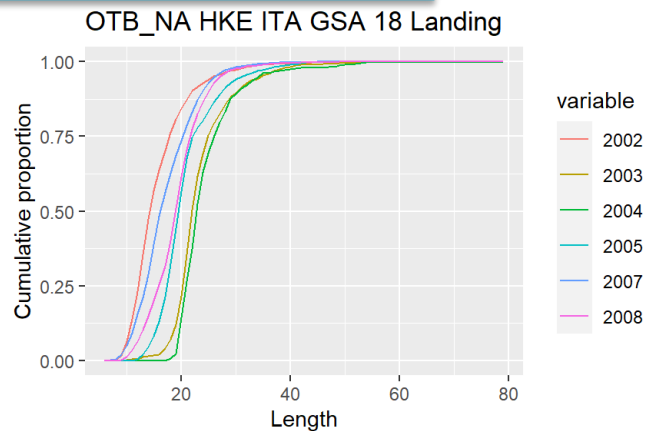
Check of SOP: check presence of critical SOP values with a percentage difference greater than 5% (in catch table)

Checks on length distributions: (in landings and discards tables).

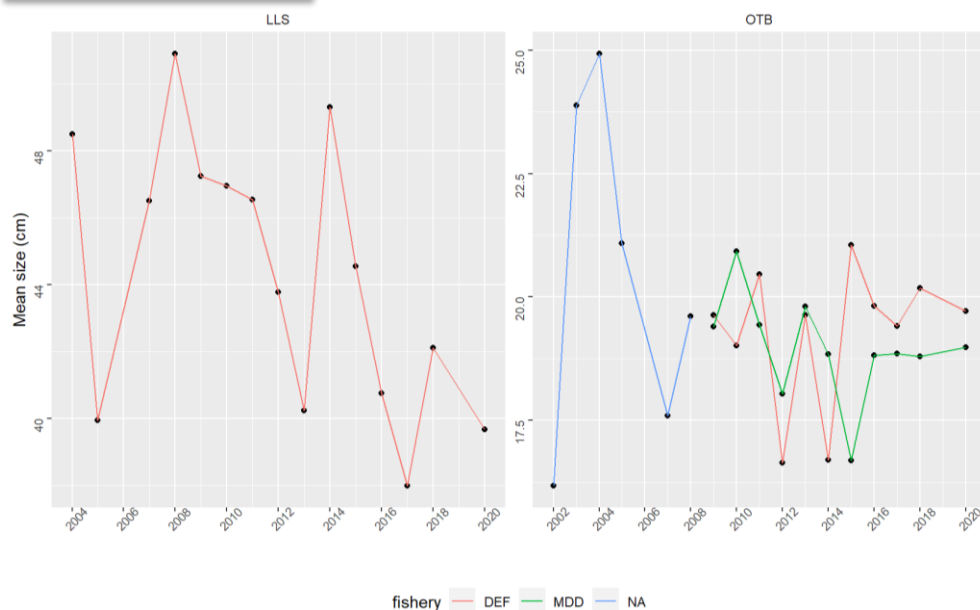
- Summary plots and tables on Length Frequency Distributions
- Check the presence of missing length classes having weights > 0
- Check the presence of weights equal to 0 having length classes filled in
- Check the presence of weights equal to -1 having length classes filled in
- Check the consistency of main length size indicators in the time series
- Check the presence of years with missing length distributions
- Check of cumulative length distributions by fishery and year to assess if they belong to the same distribution (Kolmogorov-Smirnov)

consistency of mean weights: tables and plots of mean weights by year, gear and fishery (in landings and discards tables).

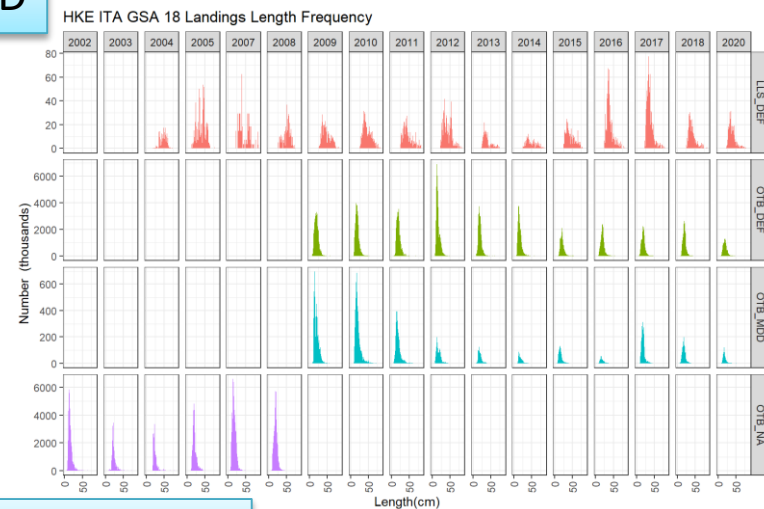
Kolmogorov-Smirnov test



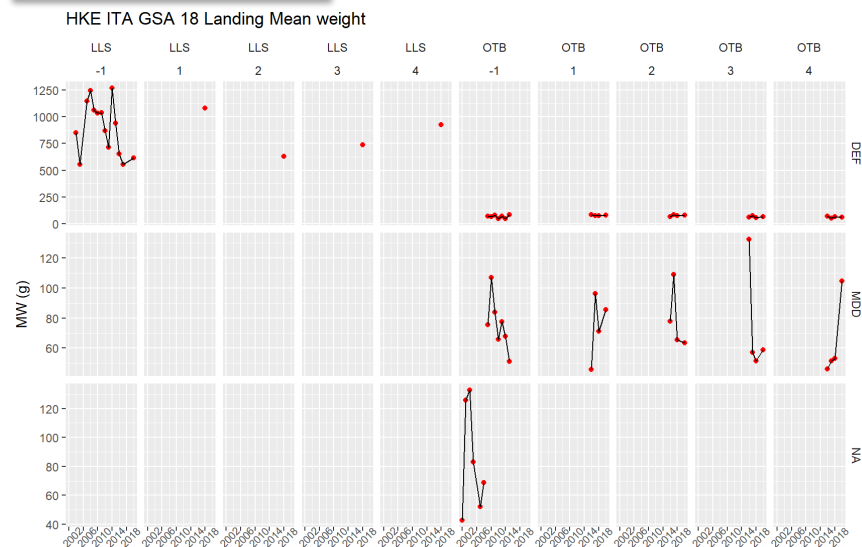
Mean lengths



LFD

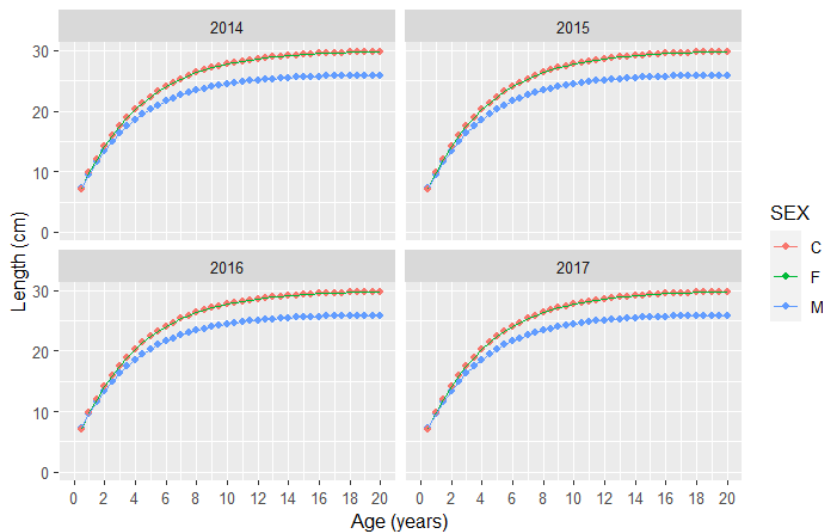


Mean weights



Growth parameters (GP) table

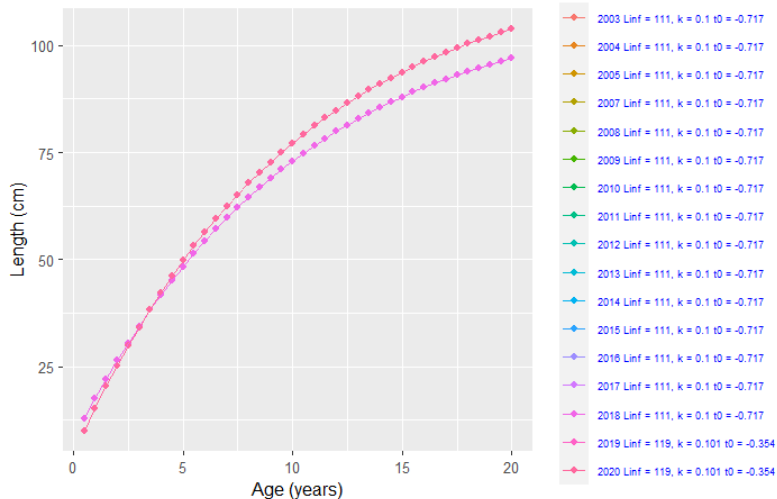
VBGF curve of MUT in ITA - GSA 18



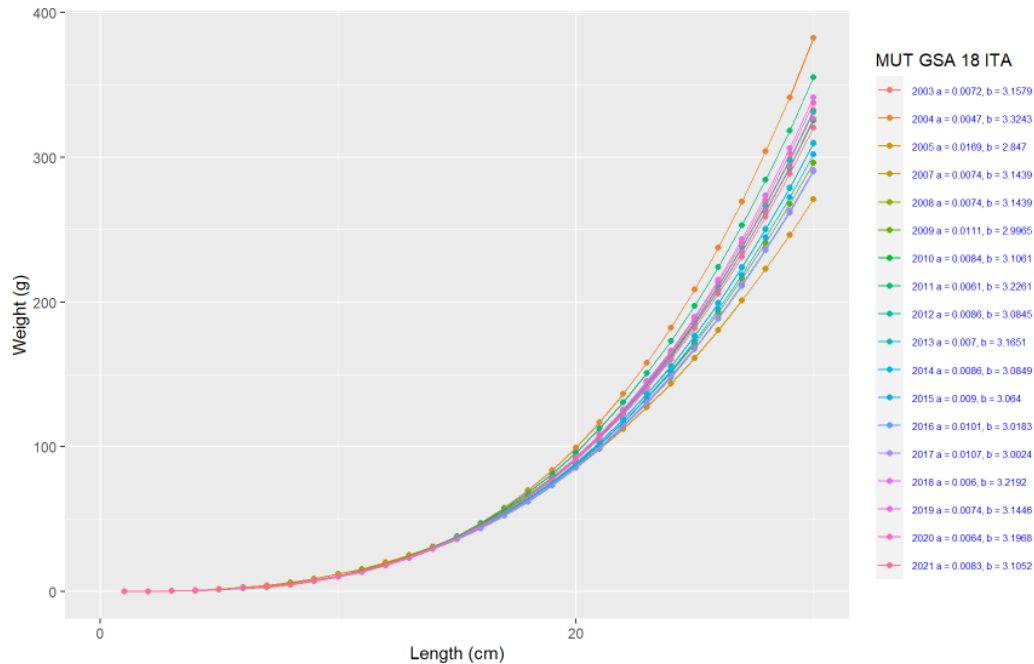
Checks on GP table:

- Plots of the time series of growth parameters by sex and year
- Plots of the time series of length-weight parameters by sex and year

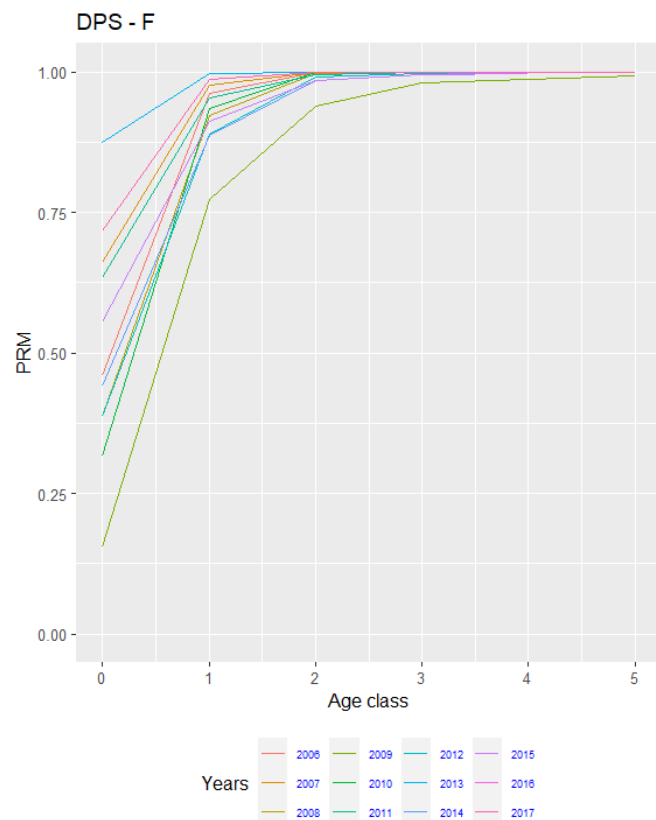
VBGF curve of F HKE in ITA - GSA 18



LW curve of M MUT in ITA - GSA 18



Maturity at Age(MA) table



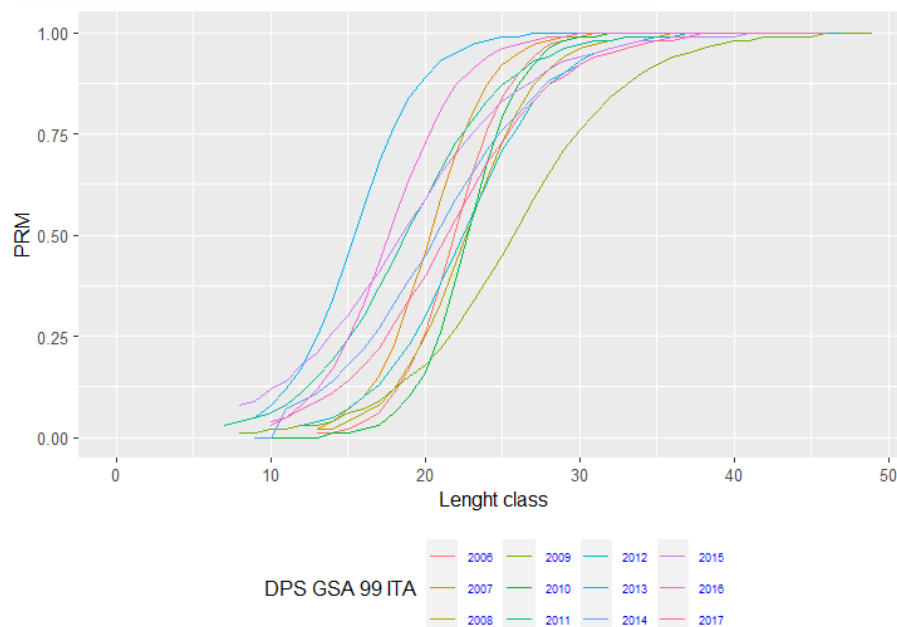
Checks on MA table:

- Plots and tables to check the of the proportion of matures for age class by sex and year

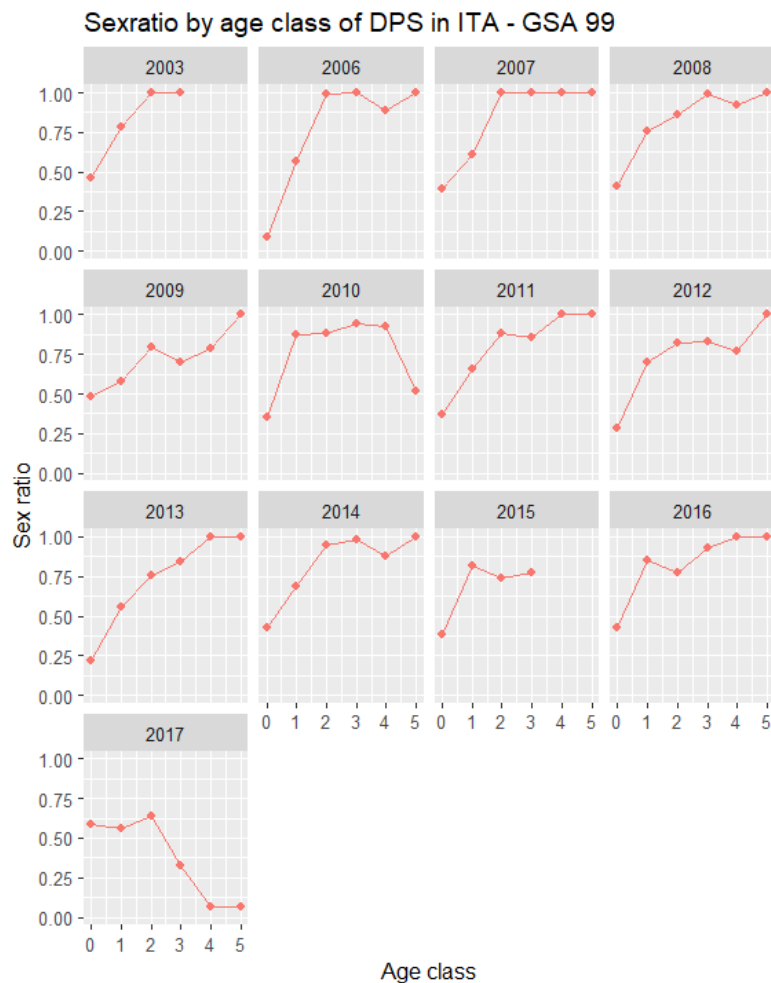
Checks on ML table:

- Plots and tables to check the proportion of matures for length class by sex and year

Maturity at Length (ML) table



Sex Ratio at Age (SRA) table



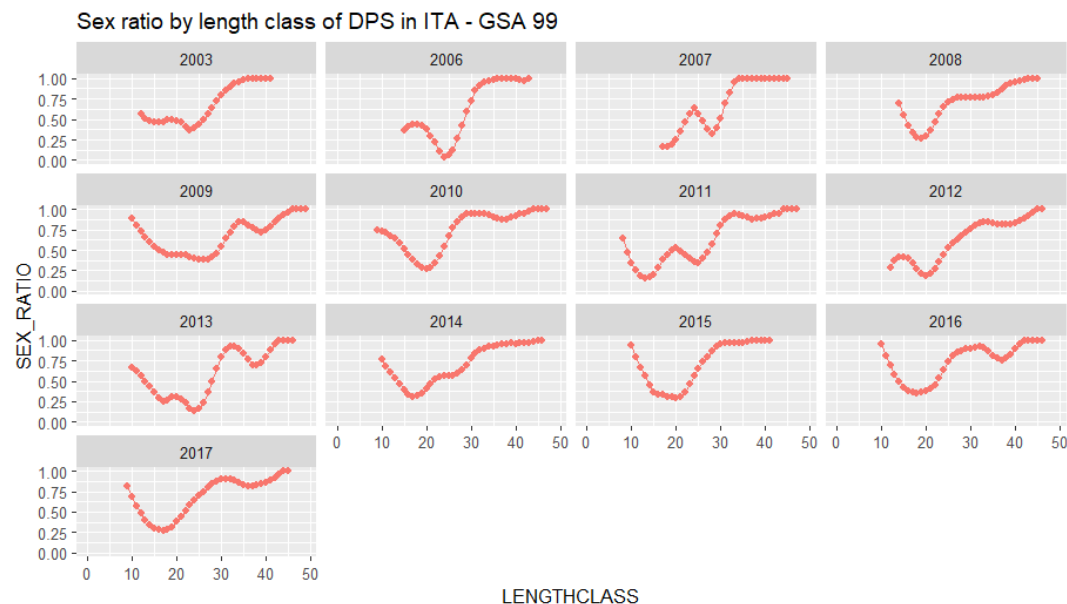
Checks on SRA table:

- Plots and tables to check the sex ratio for age class by year

Checks on SRL table:

- Plots and tables to check the sex ratio for length class by year

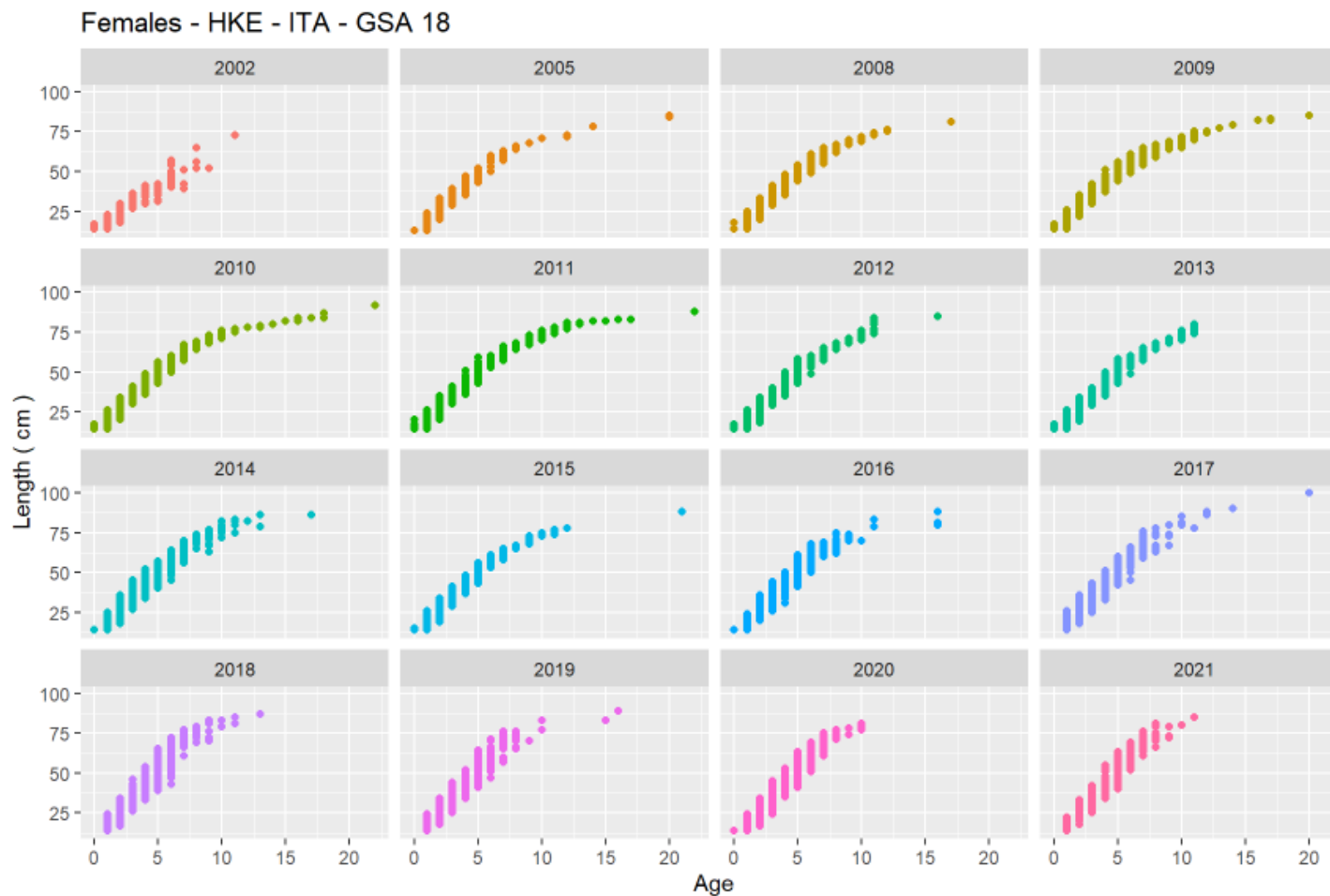
Sex Ratio at Length (SRL) table



Age-Length Keys (ALK) table

Checks on ALK table:

- Plots of Age-Length Keys



RDBqc: data quality checks for FDI format data call

The package includes functions relative to the relevant tables for Mediterranean and Black Sea.

- Table A
- Table G
- Table H
- Table I
- Table J

The functions aimed at verifying :

- the consistency of transversal variables (effort and landing);
- the spatial coverage;
- temporal coverage.

Cross-checks among the tables have been also implemented, following the checks described in the **STECF EWG 21-12 report**.

Table	Check functions	Table	Check functions
Table A	FDI_Check_empty_field_A	Table I	FDI_Check_empty_field_I
Table A	FDI_Check_record_duplicated_A	Table I	FDI_Check_record_duplicated_I
Table A	FDI_coverage	Table I	FDI_coverage
Table A	FDI_cov_tableA	Table I	FDI_checks_spatial_HI
Table A	FDI_disc_coverage	Table I	FDI_check_coord
Table A	FDI_prices_not_null	Table J	FDI_Check_empty_field_J
Table A	FDI_price_cov	Table J	FDI_Check_record_duplicated_J
Table G	FDI_Check_empty_field_G	Table J	FDI_coverage
Table G	FDI_Check_record_duplicated_G	Table J	FDI_cov_tableJ
Table G	FDI_coverage	Table J	FDI_vessel_lenth
Table G	FDI_cov_tableG	Cross checks	FDI_fishdays_cov
Table H	FDI_Check_empty_field_H	Cross checks	FDI_landweight_cov
Table H	FDI_Check_record_duplicated_H	Cross checks	FDI_vessel_numbers
Table H	FDI_coverage	Cross checks	FDI_Cross_checks_AG
Table H	FDI_checks_spatial_HI	Cross checks	FDI_Cross_Checks_AH
Table H	FDI_check_coord	Cross checks	FDI_Cross_Checks_IG

RDBqc: data quality checks for FDI format data call

Check of empty fields: checks the presence of not allowed empty data (in tables A, G, H, I, J)

Check of duplicated records: checks the presence of duplicated records (in tables A, G, H, I, J)

Checks on coverage: checks number of records and time series of relevant variables by country, GSA and year (in tables A, G, H, I, J) for different levels of aggregation

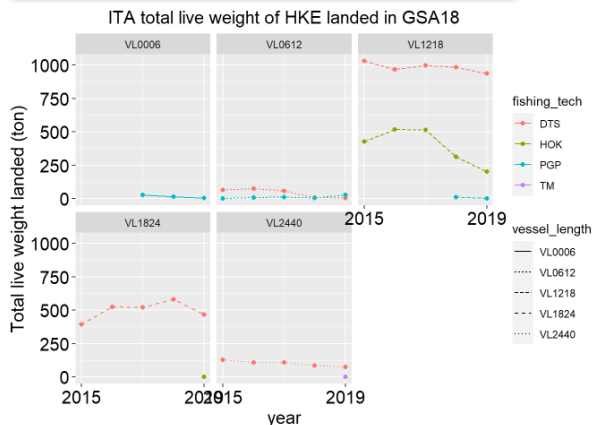
- total landing, total value, total discard (Table A)
- Totseadays, totfishdays, totkwdaysatsea, totgtdaysatsea, totkwfishdays, totgtfishdays, hrsea, kwhrsea (Table G)
- total trips, total kW, total GT, total vessels (Table J)

Consistency of prices and values: check the average price per species and the consistency between landings and landing values

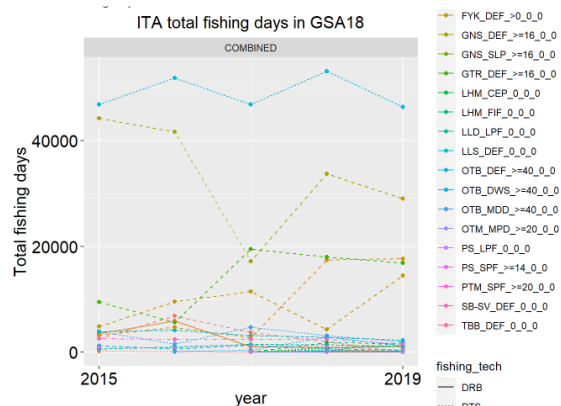
Consistency of spatial data: checks the incorrect combination of NA in the spatial columns (in table H and table I) and the compatibility of the geographical coordinates with the value provided for rectangle type.

Consistency of vessel length information: checks the average vessels length with the vessel length category (table J)

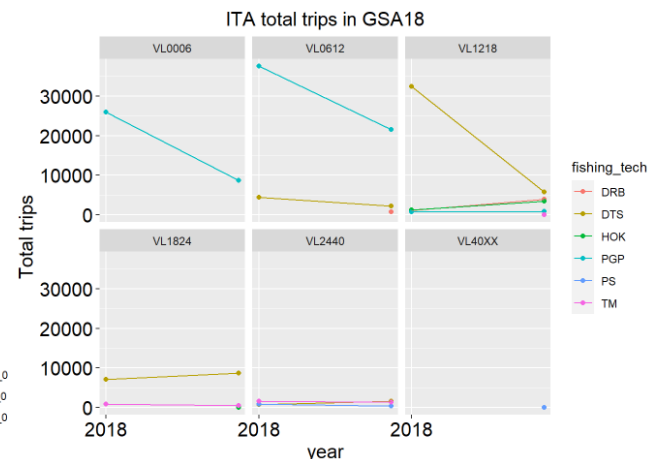
Total landing weight (A)



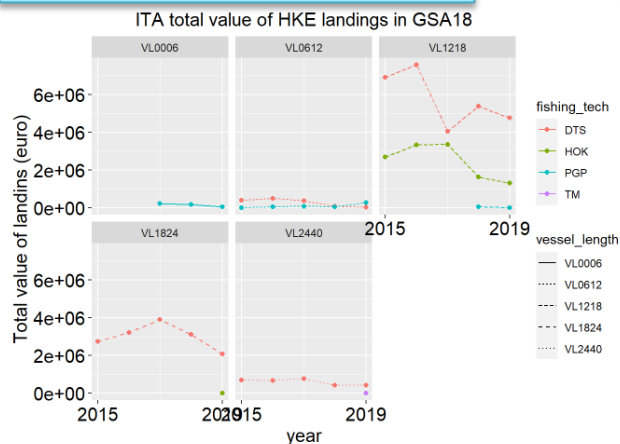
Total fishing days (G)



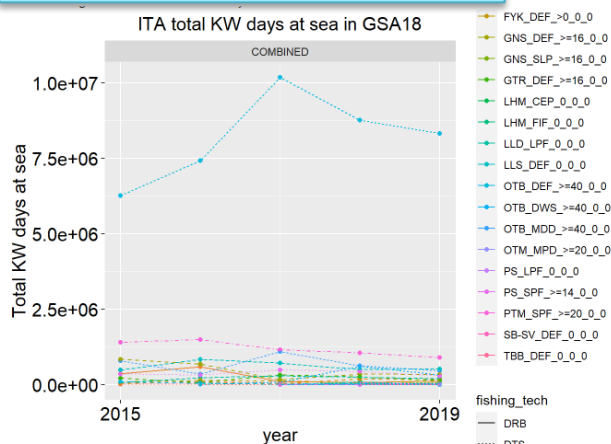
Total trips (J)



Total landing value (A)



Total kW days at sea (G)



Cross-table checks

Consistency check of:

- total fishing days in tables G and I
- landing volumes in tables A and H
- number of vessels in FDI table G and J
- landings and effort in tables A and G
- landings and spatial landings in table A and H
- effort and spatial effort in tables G and I

5.5 Consistency of vessel length with vessel length category

The FDI Capacity and fleet segment effort table (Table J) was analyzed to identify possible inconsistencies between average vessels length and vessel length category. Two tables are generated. The first one reports the records in which either vessel length or vessel category or both are not present, if any. The second table reports the cases in which vessel length does not match vessel length category, if any. [1] "Vessel length within vessel length category for all records"

Summary table of records in which either vessel lengths or vessel categories or both are not reported.

	country	year	vessel_length	principal_sub_region	totkw	totgt	totves	avgage	avgloa	consistent
584	ITA	2019	VL40XX	GSA18	412	240	1	18	46.81	NA

- No cases in which vessel length does not match vessel length category.

5.6 Comparison of total fishing days in tables G and I

Both FDI table G (Effort by rectangle) and I (Effort by rectangle) were analysed to report the mutual consistency of the total fishing days and the percentage difference between the two tables.'

Summary table of total fishing days by country, GSA and year (Table G and I).

year	country	gsa	Total fish days table G	Total fish days table I	% diff
2015	ITA	GSA18	127751.6	NA	NA
2016	ITA	GSA18	136612.2	NA	NA
2017	ITA	GSA18	117258.3	9262	-92.10%
2018	ITA	GSA18	146804.0	11725	-92.01%
2019	ITA	GSA18	141876.2	NA	NA

5.7 Comparison of landing volumes in tables A and H

Both FDI table A (Catch) and H (Landings by rectangle) were analysed to report the mutual consistency of the landing volumes and the percentage difference between the two tables.'

Summary table of landing volumes by country, GSA and year (Table A and H).

year	country	gsa	Lands table A	Lands table H	% diff
2015	ITA	GSA18	4279.342	NA	NA
2016	ITA	GSA18	4638.993	NA	NA
2017	ITA	GSA18	3957.421	NA	NA
2018	ITA	GSA18	4958.911	18.834	-99.62%

RDBqc: data quality checks for GFCM DCRF data call

Specific **consistency quality checks** are available for biological information stored in:

TASK II.2 - CATCH DATA PER SPECIES

TASK III - INCIDENTAL CATCH

TASK VII.2 - LENGTH DATA

TASK VII.3.1 - SIZE AT FIRST MATURITY

TASK VII.3.2 - MATURITY DATA

<https://www.fao.org/gfcm/data/dcrf/es/>



Table	Check functions
task II.2	GFCM_Check_empty_field_II2.R
task II.2	GFCM_Check_record_duplicated_II2.R
task II.2	GFCM_cov_task2.2
task III	GFCM_Check_empty_field_III.R
task III	GFCM_Check_record_duplicated_III.R
task III	GFCM_cov_task3
task VII.2	GFCM_Check_empty_field_VII2.R
task VII.2	GFCM_Check_record_duplicated_VII2.R
task VII.2	GFCM_relationship_length_weight_VII2.R
task VII.2	GFCM_Check_lfd_VII2.R
task VII.3.1	GFCM_Check_empty_field_VII31.R
task VII.3.1	GFCM_Check_record_duplicated_VII31.R
task VII.3.1	GFCM_Check_L50_VII31.R
task VII.3.2	GFCM_Check_empty_field_VII32.R
task VII.3.2	GFCM_Check_record_duplicated_VII32.R
task VII.3.2	GFCM_check_species_catfau_VII32.R
task VII.3.2	GFCM_relationship_length_weight_VII32.R

The functions work using:

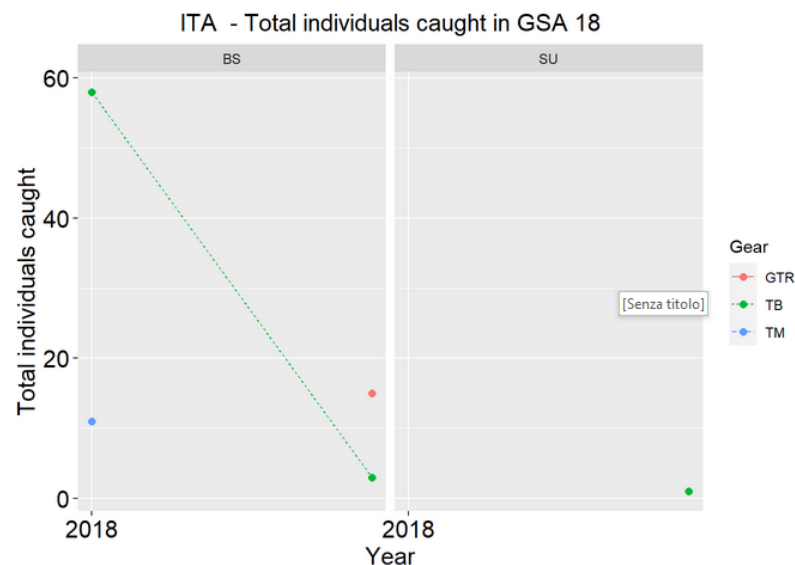
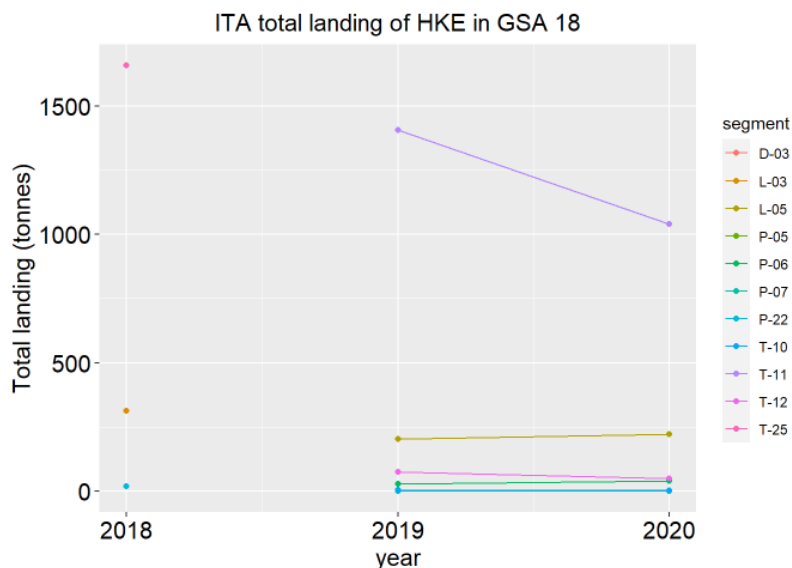
- common filters: GSA, MS, SP
- verbose parameter

Check of duplicates: covering all the tables

Check of empty fields: checks the presence of not allowed empty data (all the tables)

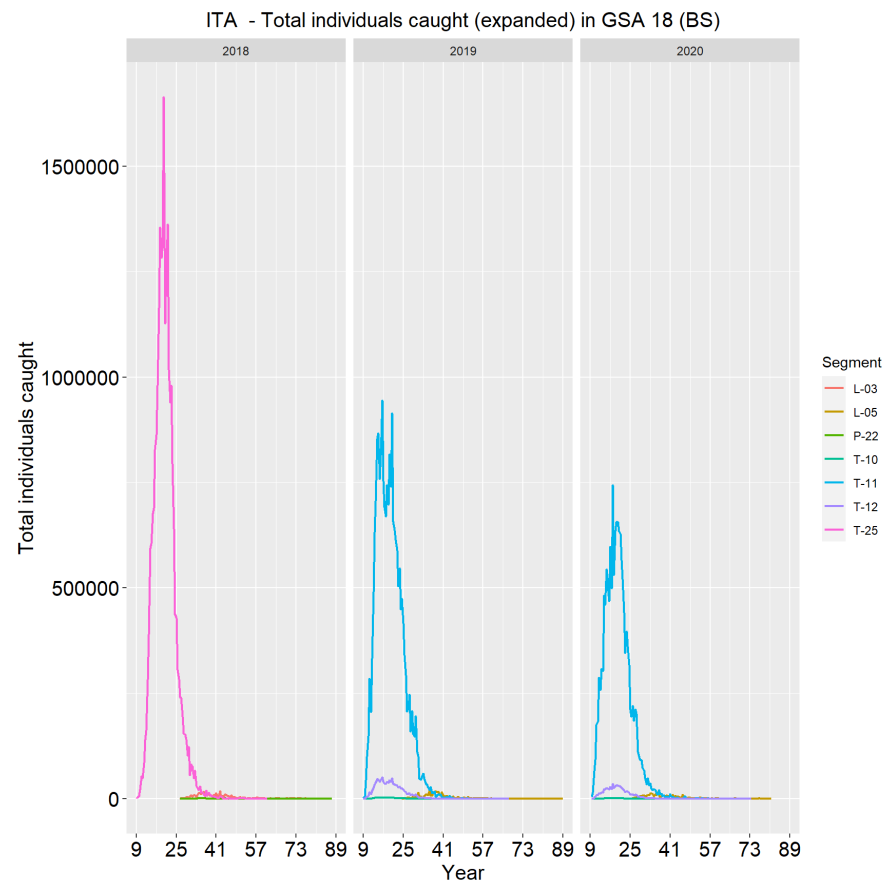
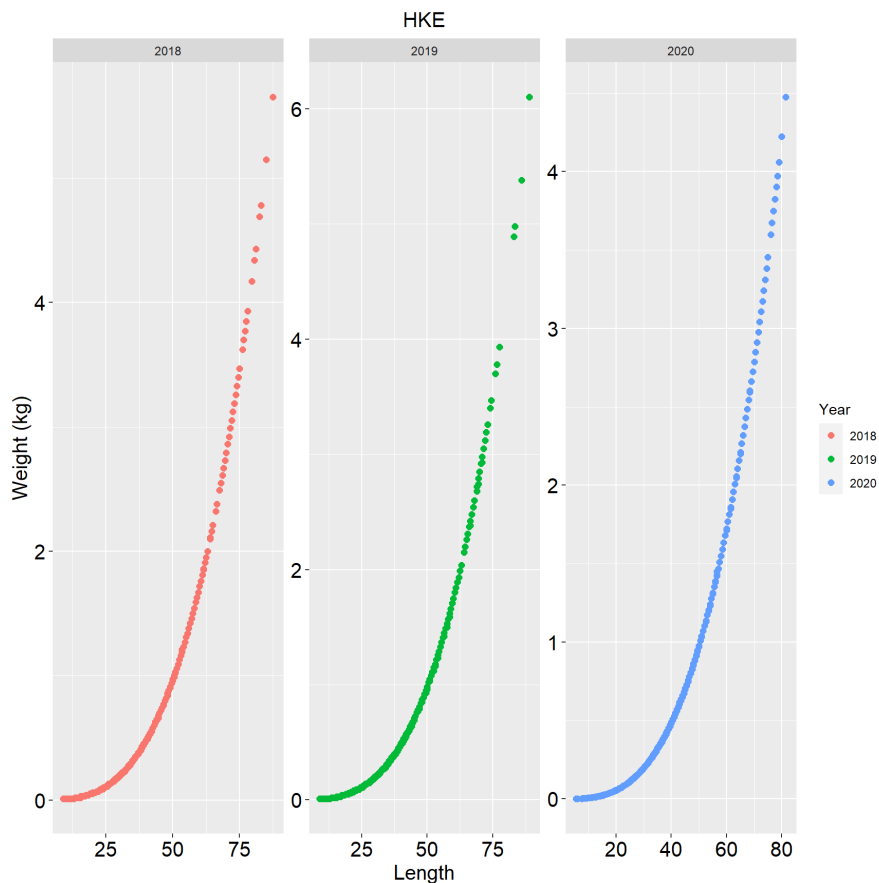
Checks on coverage: plot and table to summarise number of records and time series

- total landing and total discard (Task II.2)
- number of individuals caught (Task III)



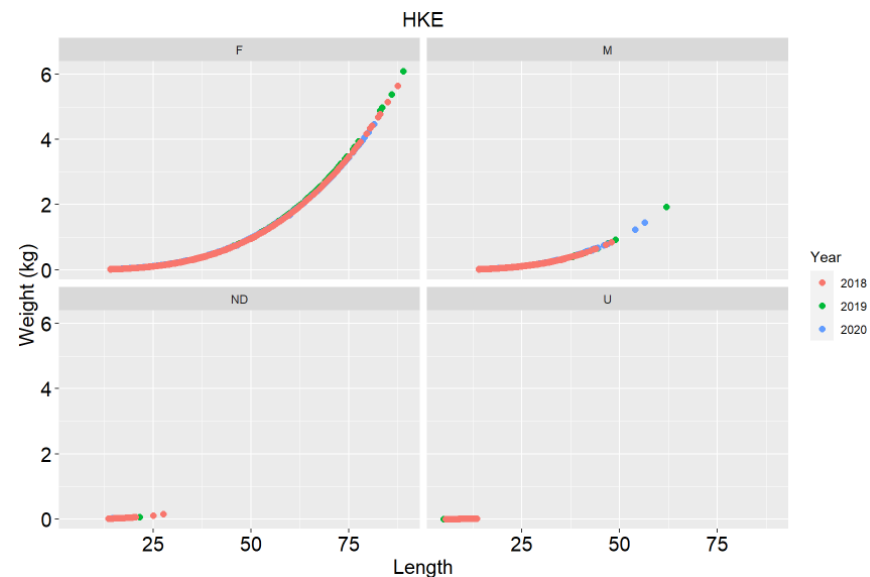
Check of biological data:

- length-weight relationships by year
- LFDs of the selected species by year, segment and source



Check of biological data:

- size at first maturity time series by sex (Table VII.3.1)
- length distribution by sex, year and maturity stage (Table VII.3.2)
- length-weight relationship by sex and year (Table VII.3.2)



Cross-checks among Data Calls

Specific functions have been developed in **QualiTrain** project to **cross-checks** data reported in **different data calls**:

MED & BS vs FDI

- Comparison of annual weights (landings and discards) at country and GSA level;
- Comparison of annual weights (landings and discards) at country, GSA and gear level;
- Comparison of weights (landings and discards) at country, GSA and gear level by quarter
- Comparison of weights (landings and discards) at country, GSA and metier level

MED & BS vs FDI vs AER

- Comparison of annual weights (landings and discards) at country and GSA level;

FDI vs AER (Annual Economic Report)

- Comparison of landings and landings' values by species and year at different levels of aggregation

MED & BS vs AR (Annual Report)

- Comparison of number of age measurements
- Comparison of number of length measurements
- Comparison of number of maturity measurements
- Comparison of number of trips reported
- Comparison of number of weight measurements

RDBqc functions were organized in a structured way in R Markdown documents to generate automatic reports



MEDBS datacall Report

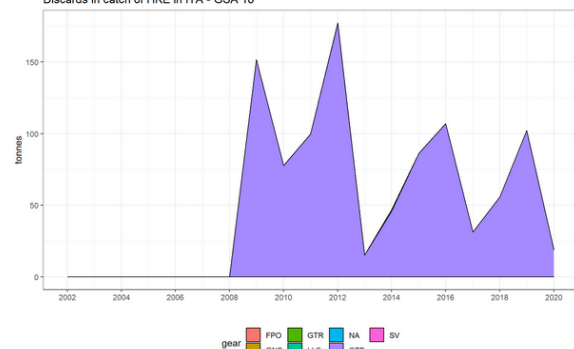
Compiled on 13/10/2022, 14:41

- 1 Catch table
 - 1.1 Presence of duplicat
 - 1.2 Coverage of catch d
 - 1.2.1 GSA 18
 - 1.2.1.1 HKE
 - 1.2.2 GSA 19
 - 1.2.2.1 HKE
 - 1.3 Check of Sum of Pro
 - 1.3.1 GSA 18
 - 1.3.1.1 HKE
 - 1.3.2 GSA 19
 - 1.3.2.1 HKE
- 2 Landing table
 - 2.1 Presence of duplicat
 - 2.2 Coverage of landing
 - 2.2.1 GSA 18
 - 2.2.1.1 HKE
 - 2.2.2 GSA 19
 - 2.2.2.1 HKE
 - 2.3 Checks on landing v
 - 2.3.1 Comparison

Coverage of discard data for HKE in GSA 18

country	year	area	species	discards
ITA	2009	GSA 18	HKE	151.60242
ITA	2010	GSA 18	HKE	77.86157
ITA	2011	GSA 18	HKE	99.69113
ITA	2012	GSA 18	HKE	177.32864
ITA	2013	GSA 18	HKE	15.36051
ITA	2014	GSA 18	HKE	45.79376
ITA	2015	GSA 18	HKE	86.24911
ITA	2016	GSA 18	HKE	107.10148
ITA	2017	GSA 18	HKE	31.43287
ITA	2018	GSA 18	HKE	85.88577
ITA	2019	GSA 18	HKE	102.13291
ITA	2020	GSA 18	HKE	18.97022

Discards in catch of HKE in ITA - GSA 18



R Markdown files characteristics:

Are able to work:

- on a single table
- on a group of tables
- creation of a table of contents with direct hyperlinks to reach the selected section of the report
- automatic detection / user definition of the reference Member State, areas and species on which to perform the analysis
- Use all QC functions included in the RDBqc R package, generating summary tables and plots useful to check data quality



Automatic Reporting Facilities



4 different R markdown files for automatic reporting have been developed in RDBFIS project. MED&BS Rmd file has been consolidated in the frame of QualiTrain project:

- Rmd R Markdown report file for the *a priori* QC on the RCG format,
- Rmd R Markdown report file for the *a posteriori* QC on DGMARE Med&BS format,
- Rmd R Markdown report file for the *a posteriori* QC on FDI format,
- Rmd R Markdown report file for the *a posteriori* QC on GFCM DCRF format.



Rmd files available on the following GitHub repository:

<https://github.com/COISPA/RDBqc/tree/main/RMD%20reports>