

ANNEX I

Guidelines in using R scripts developed for the STECF EWG 21-02

“Methods for supporting stock assessment in the Mediterranean”

Author: Alessandro Mannini

Alessandro, Mannini, Ph.D.
EC Joint Research Centre (JRC)
Directorate D - Sustainable Resources
Unit D.02 Water and Marine Resources
Via Enrico Fermi 2749
21027 Ispra (Va)
Italy
Tel. +39 0332 78 5784
e-mail: alessandro.mannini@ec.europa.eu

Table of Contents

I) Check landings / discards scripts.....	4
Data need.....	4
General structure.....	4
Stock selection.....	4
Checks lists.....	4
1) Looking for duplicated rows.....	4
2) Looking if any shifting in the length distribution was applied.....	5
3) Looking for landings in weight “NULL” (e.g. zero or -1) having some length classes with number!!.....	5
4) Looking for landings in weight more than zero but with no abundance in number.....	5
5) Computing landings in weight values by year, gear and fishery.....	6
6) Looking for “double” landings reported.....	7
7) Computing mean weight.....	7
II) Cumulative script.....	9
Data need.....	9
General structure.....	9
Stock selection.....	9
Definition of splines and percentile thresholds.....	9
Splines.....	9
Percentiles.....	10
Trying to fix Kolmogorov-Smirnov issue.....	10
Main outputs.....	11
Tables.....	11
1) Main length size indicators.....	11
2) Year with missing length distributions.....	12
3) KS Test.....	12
Plots.....	13
III) Quality checks script.....	14
Data need.....	14
General structure.....	14
Stock selection.....	14
Definition of threshold to be used in checking Sum Of Product (SOP) check.....	14
Landings in weight by gear in catch at age and at length files.....	15
Discards in weight by gear in catch at age and at length files.....	15
Maturity by length and by age.....	16
Sex ratio by length and by age.....	17
von Bertalanffy Growth Functions (VBGF).....	18
Length Weight relationships (LW).....	19
Age Length Key (ALK).....	21
Sum Of Product (SOP) landings and discards.....	21
Checking VBGF units.....	22
Checking LW parameters units along the years.....	22
IV) Filling gaps in Length Frequency Distributions.....	23
Data need.....	23
General structure.....	23
Main settings.....	23
a) Filling gaps in landings data by metier.....	24
1) Original length frequency distributions.....	24

2) Available metier and metier to be reconstructed.....	25
3) Metier reconstructed.....	26
4) Final length frequency distributions.....	27
5) Rising factor (aka SOP_ratio) applied in reconstructing missing LFDs.....	28
b) Filling gaps in discards data by metier.....	29
1) Original length frequency distributions.....	29
2) Available metier and metier to be reconstructed.....	30
3) Metier reconstructed.....	31
4) Final length frequency distributions.....	32
5) Rising factor (aka SOP_ratio) applied in reconstructing missing LFDs.....	33
6) Metier mismatch.....	33
V) Relative weights.....	34
Data need.....	34
General structure.....	34
Stock selection.....	34
Acknowledgment.....	34

I) Check landings / discards scripts

Data need

DCF format landings/discards by length (see in folder data the file “**landings.csv**” or “**discards.csv**”)

General structure

A folder having two subfolders “**script**” and “**data**” have to be set.

The “**script**” subfolder must be set as working directory folder.

The outputs will be stored in a new folder (create by the code) called “**output/COMMERCIAL**”

Output/COMMERCIAL folders will have subfolders based on the selected stock:

output/COMMERCIAL/country/gsa/species/Landing

(e.g. **output/COMMERCIAL/ITA/11/ARA/Landing**)

for landings and

output/COMMERCIAL/country/gsa/species/Discard

for discards data

For all the other outputs (see below “landings” will be replaced by “discards” when discards data are analysed)

Stock selection

The code works at stock level as we need to declare for which country, gsa and species we want perform the analysis.

Choose Member state and area to check

MS <- "ITA"

GSA <- "11"

SP <- "ARA"

Selected dataset will be saved in the output folder as csv file “**landings.csv**”

Checks lists

1) Looking for duplicated rows

Checking if there are duplicated records

table(duplicated(land[,c(2:13)]))# FALSE means no duplicated records are present

row_to_del <- which(duplicated(land[,c(2:13)]))

dupl <-land[which(duplicated(land[,c(2:13)])),]

If duplicated rows will be found the two lines below help in deleting these rows from the dataset.

land=land[-c(row_to_del),]

table(duplicated(land[,c(2:13)]))

A csv file call “**landings_duplicates.csv**” will be create in which duplicated records will be stored or the file will be empty if not duplicated rows will be spotted.

2) Looking if any shifting in the length distribution was applied

Check specon values to detected shifting in length distribution
unique(land\$specon)

specom is a free text field in the dataset in which could be reported additional information about the dataset (for example no length sample available, etc). In some cases has been observed that in this filed were reported information on the “true” starting point of the length distribution provided ending up in values referring to the wrong measures.

In the script no correction has been implemented because length shifting is a violation of the DCF template format and should be reported as a data transmission issue by the DTMT online tool.

3) Looking for landings in weight “NULL” (e.g. zero or -1) having some length classes with number!!

Check if there are landings in weight equal to zero or -1 having length class filled in

a) If there are any records having abundance in number and weight equal zero the code will print some warning messages while “suspect” data will be stored in the "**zero_land_vs_cl.csv**" file otherwise the message will be "**There arent zero landings**"

b) If there are any records having abundance in number but weight equal to -1 the code will print a some warning messages while “suspect” data will be stored in the "**minus1_land_vs_cl.csv**" file otherwise the message will be "**There arent -1 landings**"

4) Looking for landings in weight more than zero but with no abundance in number

Check if there are landings in weight >0 having length class not filled in

This quality checks will illustrate by printed messages about number of records with landings in weight more than zero which have or not length frequency distribution filled in (missing LFD basically). Actually, provided landings more than zero without LFD is admitted (especially for metier not selected by ranking system and/or without biological samples). Anyway all these cases will be stored in the following files given to enduser the opportunity in checking if there coverage and /or quality issues.

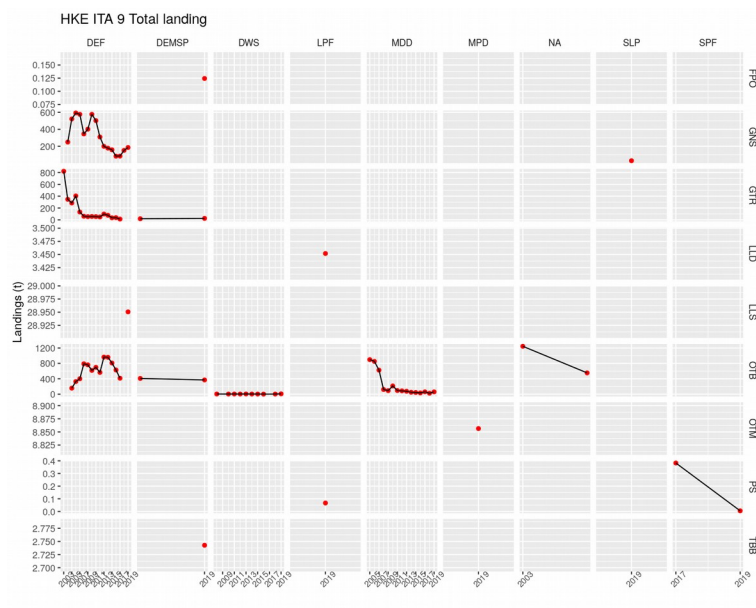
The selected records will be stored in the file “**minus2_land_vs_cl**”

5) Computing landings in weight values by year, gear and fishery

Check landings values

The values of landings in weight will be aggregated by **year**, **gear** and **fishery**.

The results will be stored in the **output** folder both as plot “**landings_weight.tiff**” and csv file “**landingsWeight.csv**”



Landings will be also plotted as coloured areas by gear and by quarter



6) Looking for “double” landings reported

a) ##### Comparison between landings in weight by quarter and -1 (year aggregation)

Landings in weight will be aggregated at year and gear level summing up available quarters and quarters reported as “-1” (usually this values stands for “year level”).

Then a ratio over the two values is computed.

Ratio equal 1 or very closed to 1 deserve a deeper check. All the other values including NA should be fine.

year	gear	tot_q	tot_yr	ratio
2005	GNS	NA	521.70634	NA
2005	GTR	NA	284.3715	NA
2005	OTB	NA	1053.90027	NA
2003	GTR	NA	820.55742	NA
2003	OTB	NA	1243.90435	NA
2004	GNS	NA	249.53592	NA
2004	GTR	NA	346.36224	NA
2004	OTB	NA	552.89795	NA
2006	GNS	NA	592.56606	NA
2006	GTR	NA	403.96055	NA

The output file is stored as: "**Comparison_landing_quarter_-1.csv**"

b) ## Comparison between landings in weight by quarter and -1 (year aggregation) assign to different Vessel Length

This one is an additional check to look for is similar issues spotted by the EWG2015 in some areas are still reported in the dataset . In this case the landings were reported twice by year, gear and quarter but in one case the vessel length was actually define while in the other one was “NA”

year	gear	quarter	tot_VL	tot_NoVL	ratio
2005	GNS	-1	NA	521.70634	NA
2005	GTR	-1	NA	284.3715	NA
2005	OTB	-1	NA	1053.90027	NA
2003	GTR	-1	NA	820.55742	NA
2003	OTB	-1	NA	1243.90435	NA
2004	GNS	-1	NA	249.53592	NA
2004	GTR	-1	NA	346.36224	NA

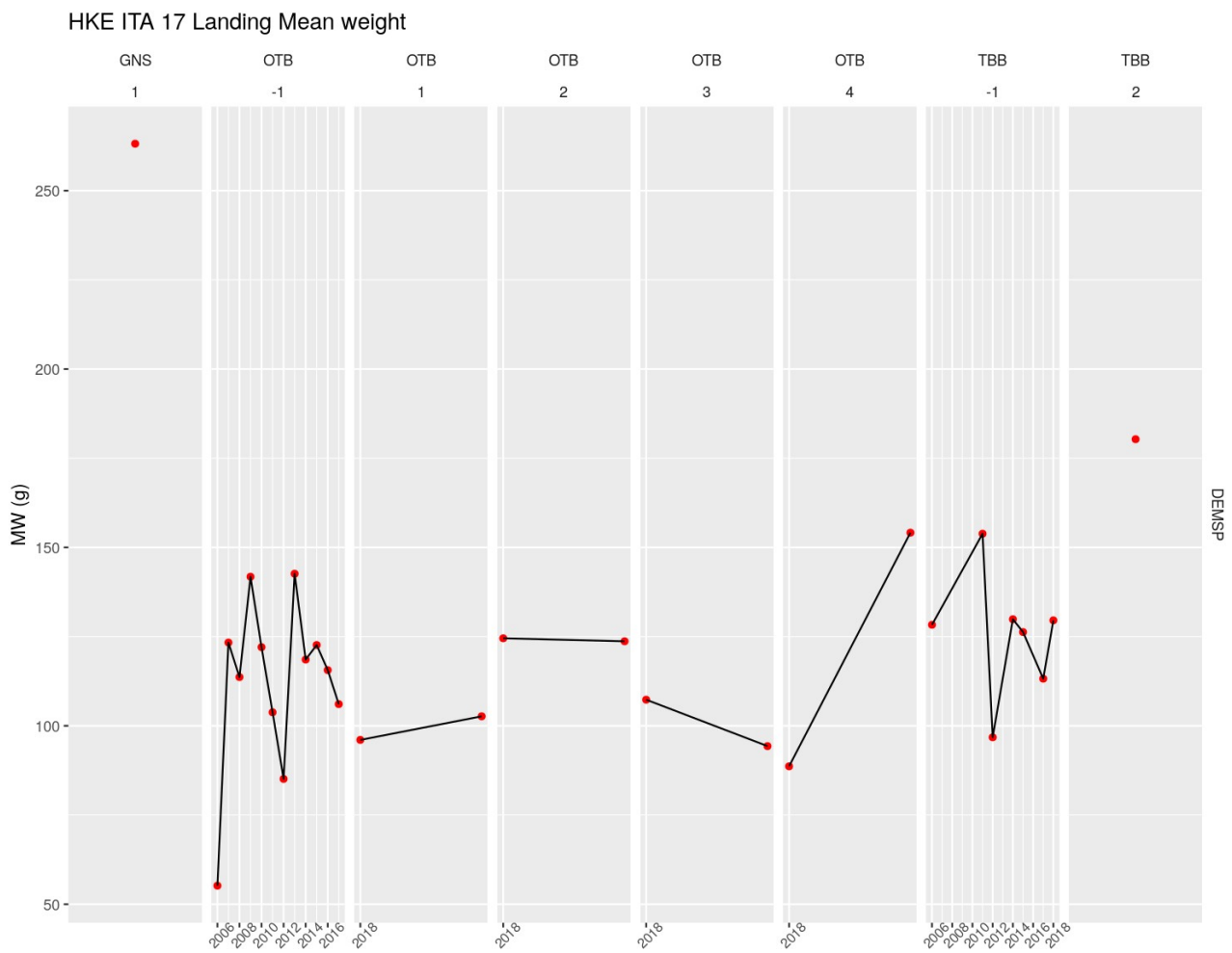
The output file is stored as: "**Comparison_landing_quarters_VL.csv**"

7) Computing mean weight

Mean Weight check

Mean weight by year, gear, fishery and quarter aggregation is computed.

Values are stored in the "**MW_land.csv**" while plots are saved as "**landings_mean_weight.tiff**"



Suspicious values need to be checked exploring the dataset trying to understand if the values are due to coverage (e.g. landings for 4 quarters but number only for some of them) and/or quality issues (e.g. unit of measures misreported).

II) Cumulative script

Data need

DCF format landings/discards by length (see in folder data the file “**landings.csv**” or “**discards.csv**”)

General structure

A folder having two subfolders “**script**” and “**data**” have to be set.

The “**script**” subfolder must be set as working directory folder.

The outputs will be stored in a new subfolder (create by the code) called “**Cumulative**” (the whole path will be: output/COMMERCIAL/country/gsa/species/Cumulative)

Stock selection

The code works at stock level as we need to declare for which country, gsa and species we want perform the analysis.

Choose Member state, area and species to check

MS <- "ITA"

GSA <- "11"

SP <- "ARA"

Selected datasets will be saved in the output folder as csv file “**LFLandings.csv**” and “**LFDiscards.csv**”

Definition of splines and percentile thresholds

Splines

splines <- c(0.2,0.4,0.6,0.8) ##

Spline values assignment to fit cumulative distributions.

Inverse function is used to estimate length size corresponding to the different percentiles. To save plots on how cumulative distributions are fitted by the different splines the following lines in the landings and discards loop need to be activated. Be aware that saving all the plots can takes time.

Lines 213 -230 for landings

Modelling cumulative frequencies

p1 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

geom_point()+stat_smooth(method="loess",fullrange=T,span=0.2,col=2)+ggtitle("Spline 0.2")

p2 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

```

# geom_point()+stat_smooth(method="loess",fullrange=T,span=0.4,col=2)+ggtitle("Spline
0.4")

# p3 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

# geom_point()+stat_smooth(method="loess",fullrange=T,span=0.6,col=2)+ ggtitle("Spline
0.6")

# p4 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

# geom_point()+stat_smooth(method="loess",fullrange=T,span=0.8,col=2)+ggtitle("Spline
0.8")

#ggsave(filename=paste0(dir_t,"/",MS,"/",SP,"/",paste0("Cumulative_LFD_LANDING_",i,"_",j,"_
",z,".tiff")),width = 10, height = 8, dpi = 150, units = "in",grid.arrange(p1, p2, p3,p4,ncol=2))

```

and Lines 445-459 for discards

Modelling cumulative frequencies

```

# p1 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

# geom_point()+stat_smooth(method="loess",fullrange=T,span=0.2,col=2)+ggtitle("Spline
0.2")

# p2 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

# geom_point()+stat_smooth(method="loess",fullrange=T,span=0.4,col=2)+ggtitle("Spline
0.4")

# p3 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

# geom_point()+stat_smooth(method="loess",fullrange=T,span=0.6,col=2)+ ggtitle("Spline
0.6")

# p4 <- ggplot(tempfish, aes(x=start_length, y=relative)) +

# geom_point()+stat_smooth(method="loess",fullrange=T,span=0.8,col=2)+ggtitle("Spline
0.8")

#ggsave(filename=paste0(dir_t,"/",MS,"/",SP,"/",paste0("Cumulative_LFD_DISCARD_",i,"_",j,"_
",z,".tiff")),width = 10, height = 8, dpi = 150, units = "in",grid.arrange(p1, p2, p3,p4,ncol=2))

```

Percentiles

Xtresholds = c(0.25,0.5,0.75) ##

Percentiles for which we are going to estimate for each LFD available corresponding length size based on the different splines settled.

Trying to fix Kolmogorov-Smirnov issue

Rt <- 1 ##

KS test is very sensitive to large sample size ending up in rejecting in almost all the cases the H_0 Hypothesis¹ Based on that I have introduced a ratio factor called R_t . If R_t is set equal 1 we are going to use the actual number of individuals by length classes while, setting another value, a ratio will be applied to reduce sample size (the shape of the distribution indeed will be not affected).

Main outputs

Tables

1) Main length size indicators

These lengths will be stored in the files “LFLANDINGS_DB.csv” and “LFDISCARDS_DB.csv” Just as example:

year	gear	fishery	total_number	min_size	mean_size	max_size	sd_size	spline	percentile	percentile_value
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.2	0.25	26.7611702927164
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.2	0.5	31.9560814852314
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.2	0.75	35.9479367228162
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.4	0.25	26.4215351534318
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.4	0.5	31.4087807740682
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.4	0.75	36.4651893652425
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.6	0.25	24.6931408018117
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.6	0.5	31.2424345562792
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.6	0.75	37.4741658708582
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.8	0.25	20.4547151635378
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.8	0.5	30.2319227486219
2003	GTR	DEF	2776.53642	15	32.0699999065742	74	7.51434853409285	0.8	0.75	39.2460389795577

The analysis is carried out by year, gear and fishery. For each combinations the following indicators are estimated:

total number of individuals in the sample

the minimum size observed

the mean size

the maximum size observed

the standard deviation

for each splines and percentile thresholds the corresponding length size value

If there is a combination of year, gear fishery for which there isn't length distribution (total number equal to 0) NA will be reported for all the other parameters.

year	gear	fishery	total_number	min_size	mean_size	max_size	sd_size	spline	percentile	percentile_value
2004	GTR	DEF	0	NA	NA	NA	NA	NA	NA	NA

1 <https://stats.stackexchange.com/questions/333892/is-the-kolmogorov-smirnov-test-too-strict-if-the-sample-size-is-large>

2) Year with missing length distributions

year	gear	fishery	total_number_landed
2004	GTR	DEF	0
2005	GTR	DEF	0
2006	GTR	DEF	0
2007	GTR	DEF	0
2008	OTB	DWS	0
2009	GTR	DEF	0
2010	OTB	DWS	0

Basically a subset of the previous table based on total number equal to zero

3) KS Test

Group	NTot_yr1	NTot_yr2	Dmax	Dcalc	H0_p0.05	Comment	ID
2004 vs 2005	45.54845	70.577497	0.074	0.328806814750582	accepted	belong to same population	GNS_DEF
2004 vs 2006	45.54845	167.996454	0.236	0.289003997746313	accepted	belong to same population	GNS_DEF
2004 vs 2007	45.54845	161.191005	0.246	0.290302371621861	accepted	belong to same population	GNS_DEF
2004 vs 2008	45.54845	64.492819	0.262	0.33483543828855	accepted	belong to same population	GNS_DEF
2004 vs 2009	45.54845	79.374713	0.197	0.321580498225111	accepted	belong to same population	GNS_DEF
2004 vs 2010	45.54845	129.811705	0.161	0.29793246858153	accepted	belong to same population	GNS_DEF
2004 vs 2011	45.54845	111.59494	0.143	0.30418311934011	accepted	belong to same population	GNS_DEF
2004 vs 2012	45.54845	69.010521	0.104	0.330267775084297	accepted	belong to same population	GNS_DEF
2004 vs 2013	45.54845	72.90555	0.179	0.326741179035844	accepted	belong to same population	GNS_DEF
2004 vs 2014	45.54845	49.000973	0.127	0.356070825435661	accepted	belong to same population	GNS_DEF
2004 vs 2015	45.54845	33.756899	0.297	0.392897504130638	accepted	belong to same population	GNS_DEF
2004 vs 2016	45.54845	18.332246	0.134	0.478504783017196	accepted	belong to same population	GNS_DEF
2004 vs 2017	45.54845	17.899902	0.306	0.482607590551587	accepted	belong to same population	GNS_DEF
2004 vs 2018	45.54845	23.281171	0.326	0.440752293146783	accepted	belong to same population	GNS_DEF
2004 vs 2019	45.54845	24.644461	0.473	0.432609713499485	rejected	not belong to same population	GNS_DEF

In csv file “**KS_test_landing.csv**” (same for discard) will be stored the multi-pairs KS test carried out for each combinations years~ID.

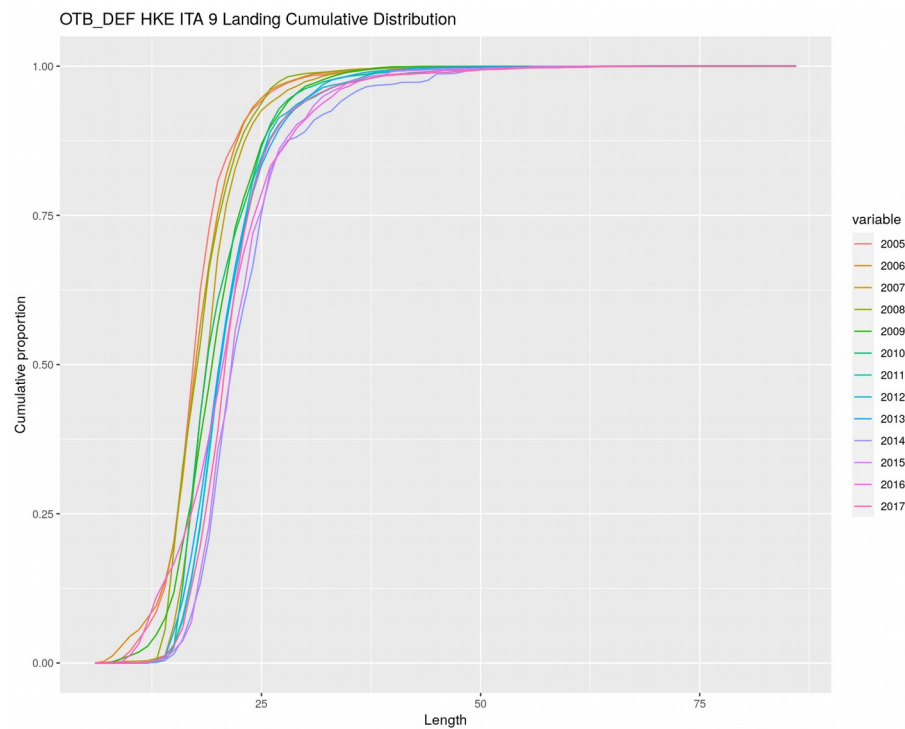
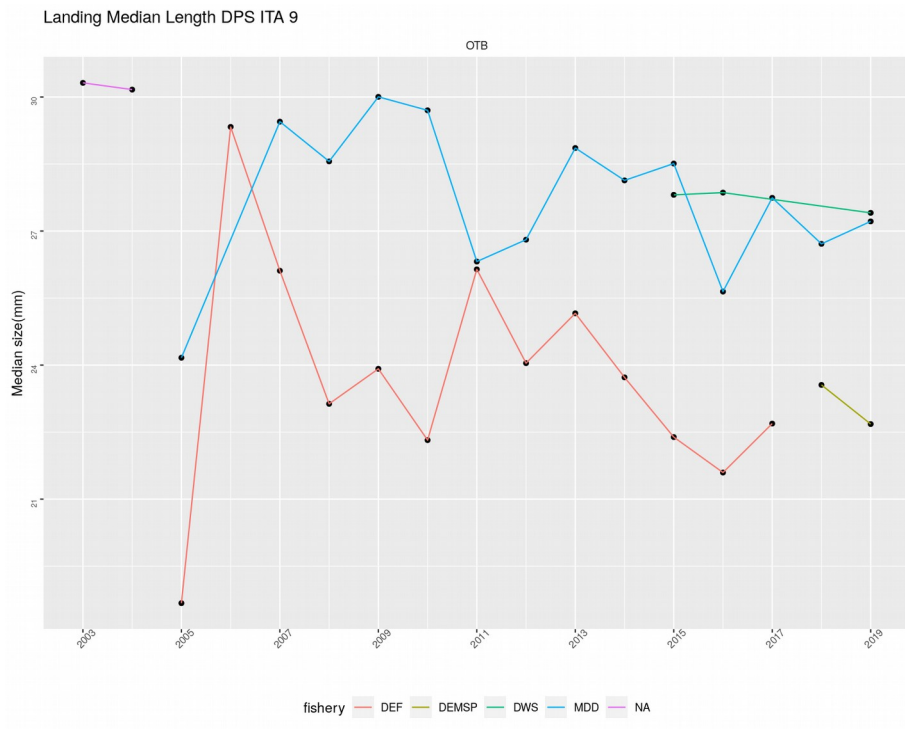
Dmax and Dcalc are the KS test values tested for a $p < 0.05$ having as H0 Hypothesis the distribution belong to the same population

If there is an ID for which is available only one LFD this LFD data will be stored in csv file called “**KS_notest_landing.csv**” (same for discards) . The file will be empty in all the other cases (namely no distributions at all or not enough distributions to run at least one comparison)

year	gear	fishery	start_length	value	ID	ID1
2013	OTB	DWS		0	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		1	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		2	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		3	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		4	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		5	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		6	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		7	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		8	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		9	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		10	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		11	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		12	0 OTB_DWS	2013OTBDWS
2013	OTB	DWS		13	0 OTB_DWS	2013OTBDWS

Plots

“Mean_LFD_LANDING.tiff” and **“Median_LFD_LANDING.tiff”** (same plots for discards)



III) Quality checks script

Data need

All DCF templates excluding surveys data and effort

General structure

A folder having two subfolders “[script](#)” and “[data](#)” have to be set.

The “[script](#)” subfolder must be set as working directory folder.

The outputs will be stored in a new subfolder (create by the code) called “[Quality_Checks](#)”

(the whole path will be: output/COMMERCIAL/country/gsa/species/QualityChecks)

Stock selection

The code works at stock level as we need to declare for which country, gsa and species we want perform the analysis.

Choose Member state, area and species to check

MS <- "ITA"

GSA <- "11"

SP <- "ARA"

Selected (subset based on the stock selection) datasets will be saved in the output folder as csv file

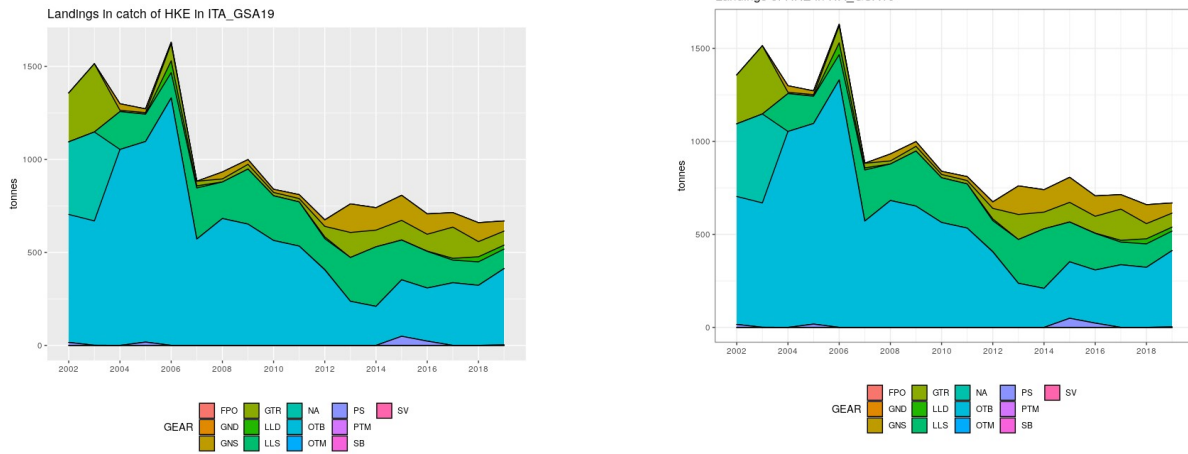
Definition of threshold to be used in checking Sum Of Product (SOP) check

threshold = 5 # SOP difference with "official" value. Only values having a SOP correction more than 1+*%threshold* (in this case 1+0.05) or less than 1-*%threshold* (1-0.05) will be saved in a csv file. The check will be based on what declared in the Catch file and will be run both for landing and discards.

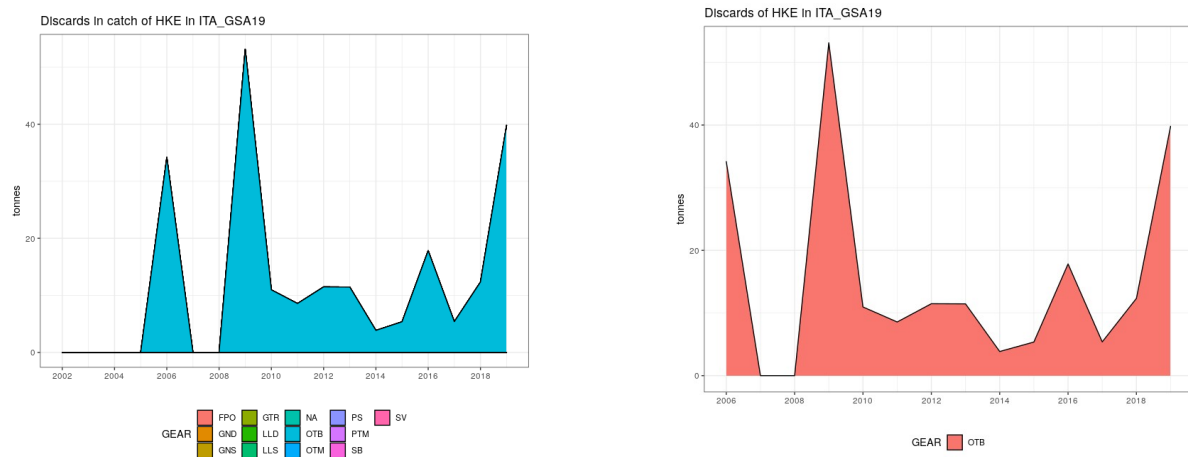
The formula to compute SOP is:

$$SOP = \sum_{i=0}^{i=\max(\text{age})} \text{number at age } i * \text{mean weight at age } i$$

Landings in weight by gear in catch at age and at length files

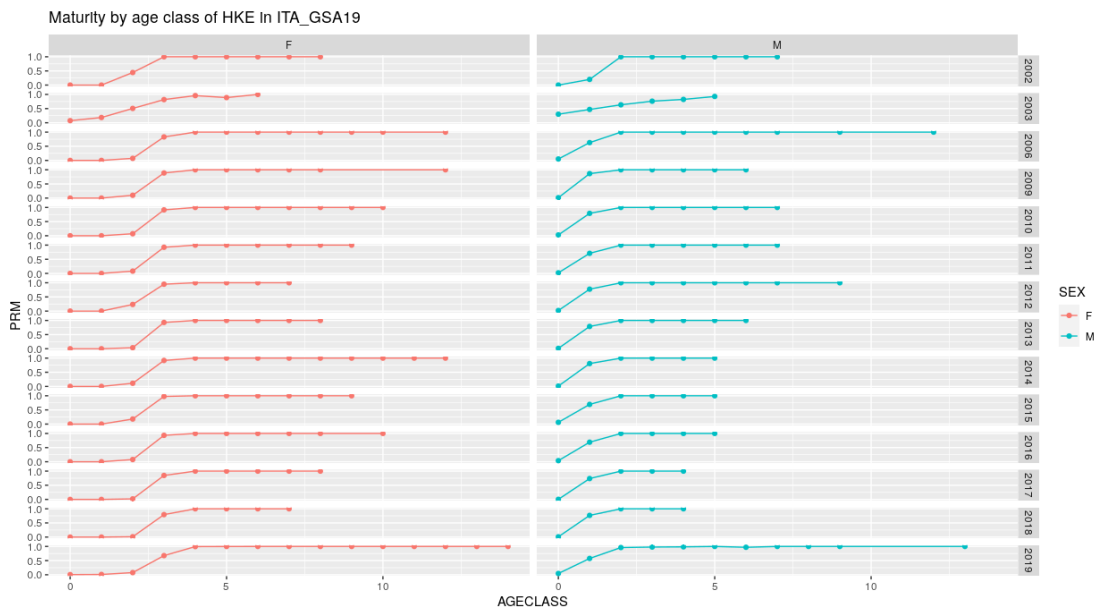
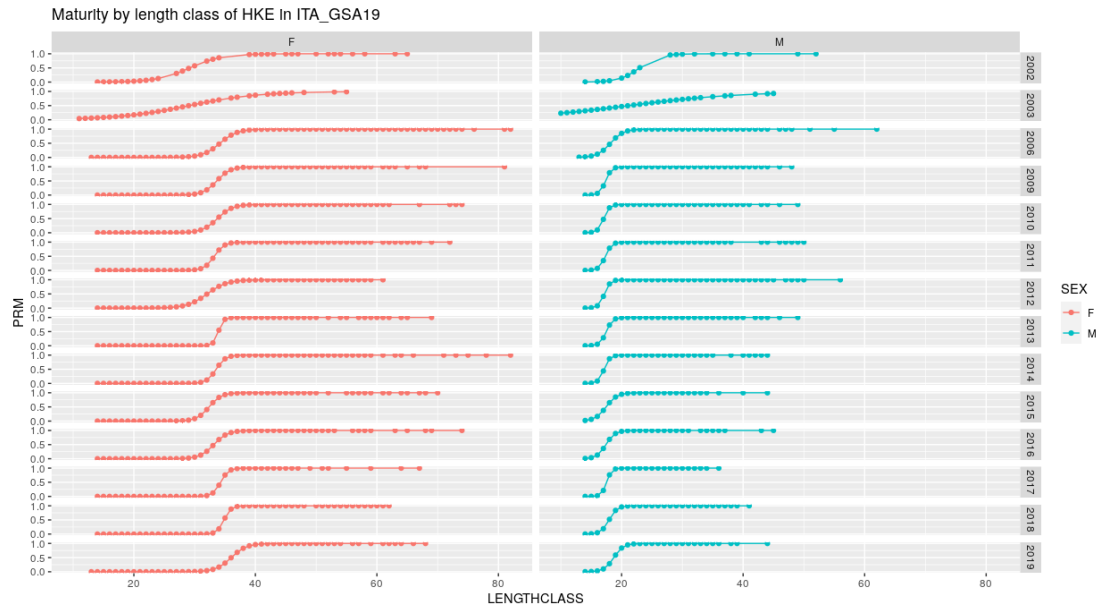


Discards in weight by gear in catch at age and at length files



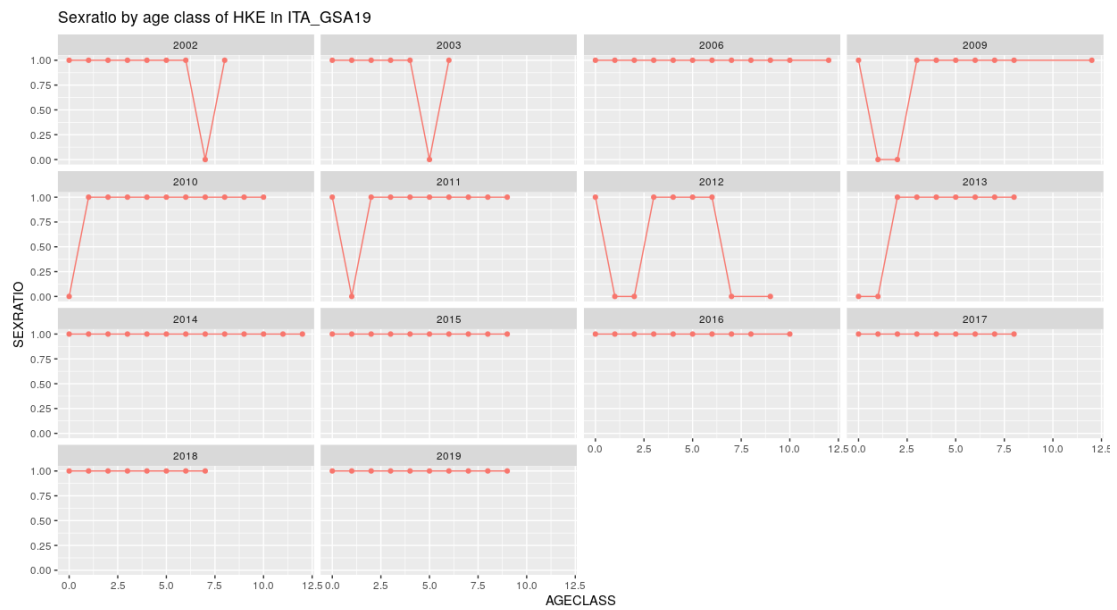
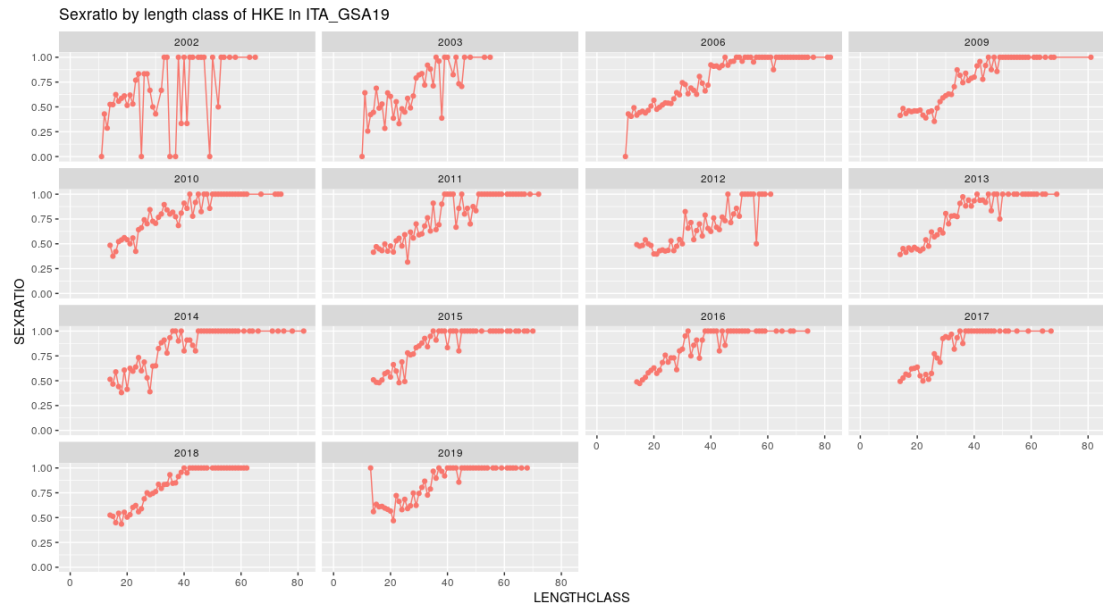
These plots (one for landings and the other ones for discards) make possible a quick comparison if the values reported in the catch at age file and landings/discards at length files are the same

Maturity by length and by age



Data extracted directly from MA and ML DCF templates will be plotted looking for trend in time and/or weird values.

Sex ratio by length and by age

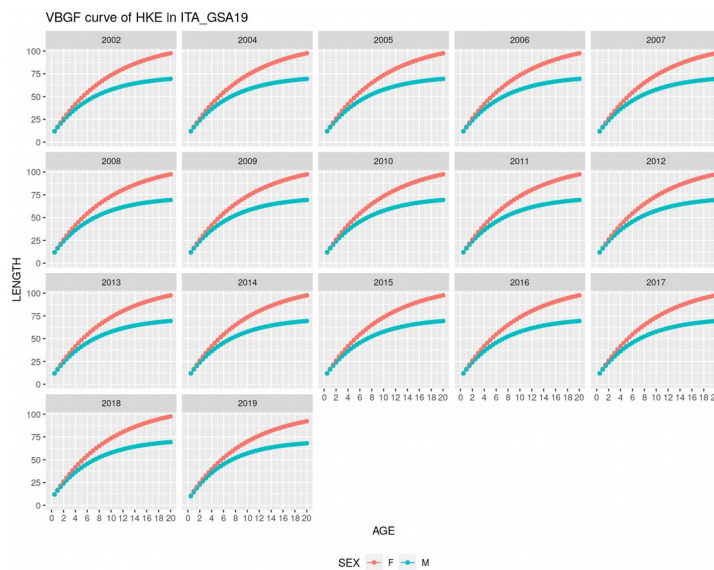


Data extracted directly from SRA and SRL DCF templates will be plotted looking for trend in time and/or weird values.

von Bertalanffy Growth Functions (VBGF)

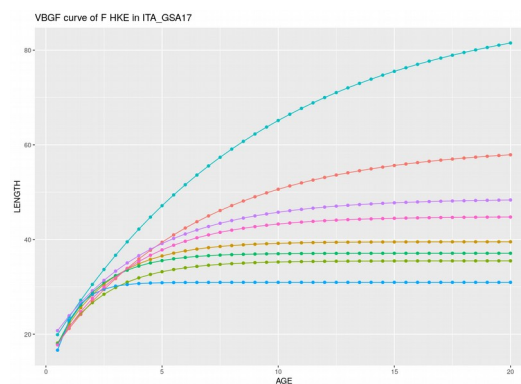
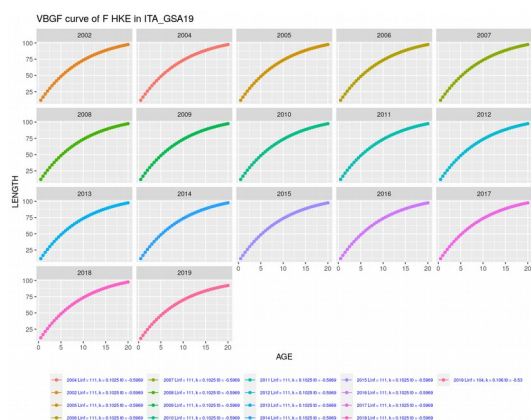
The script extracts all VBGF set available by sex and produces a series of different plots.

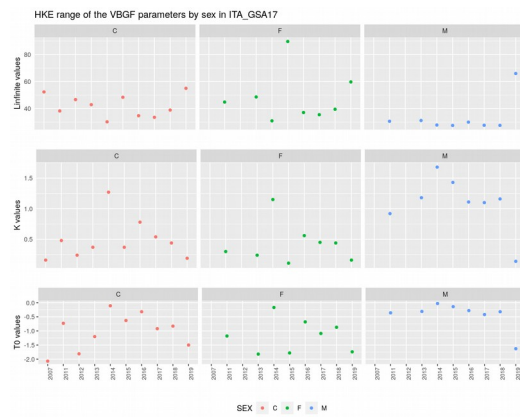
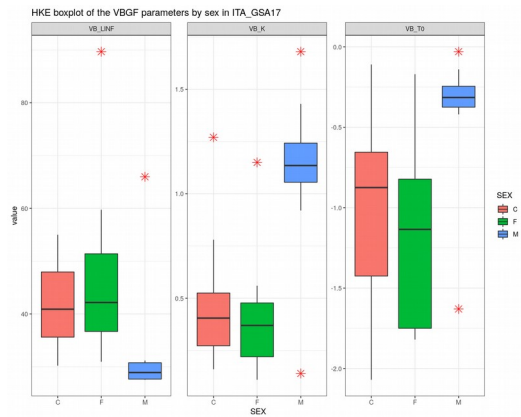
The first one is a comparison of all resulting VBGF curves by year and sex



Then for all sex available will be saved:

- plot the VBGF curves facet by year
- plot all curves in one panel
- boxplot of the VBGF parameters available
- trend in time of the VBGF parameters values

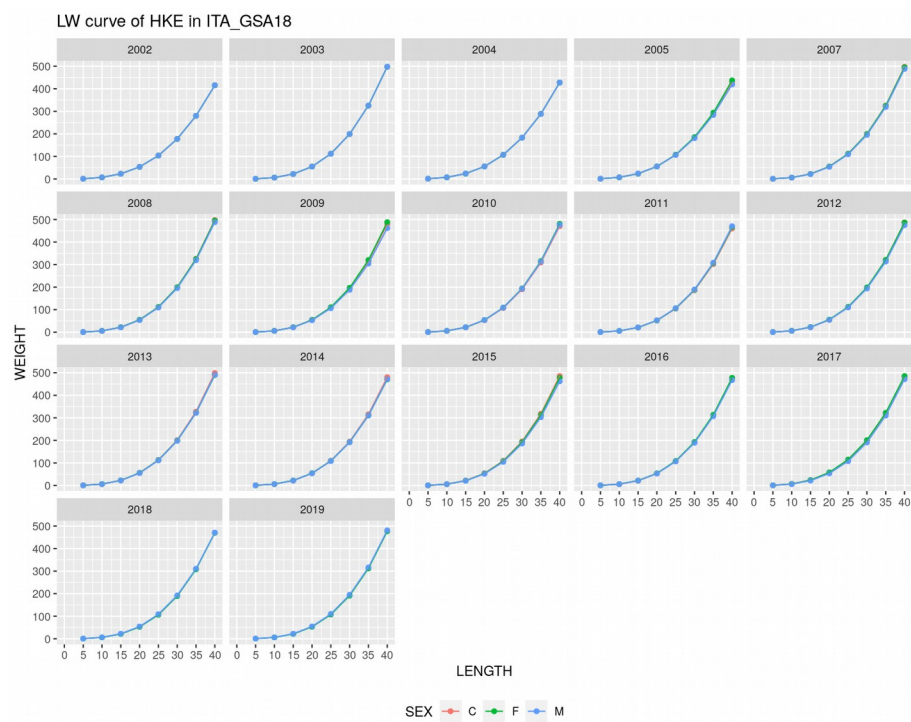




Length Weight relationships (LW)

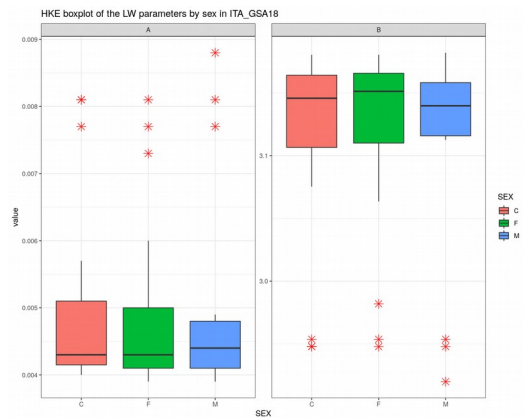
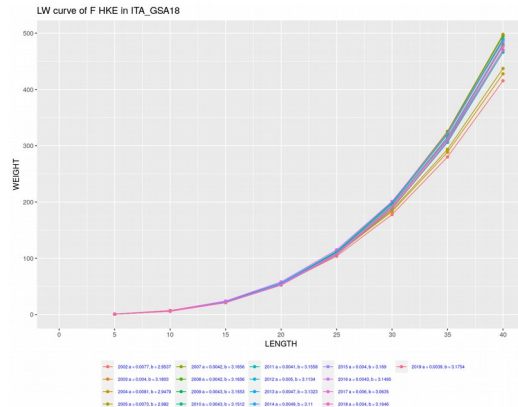
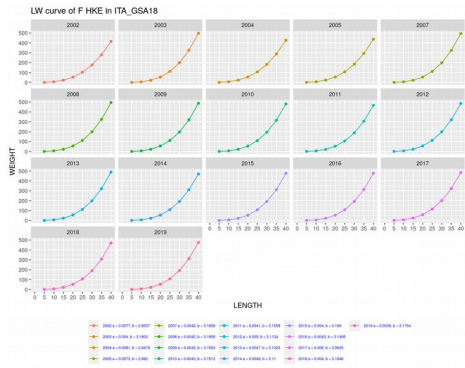
The script extracts all length weight relationships parameters available by sex and produces a series of different plots.

The first one is a comparison of all resulting LW curves by year and sex



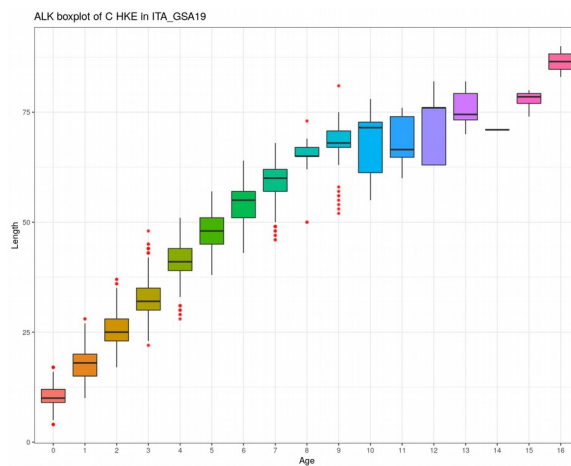
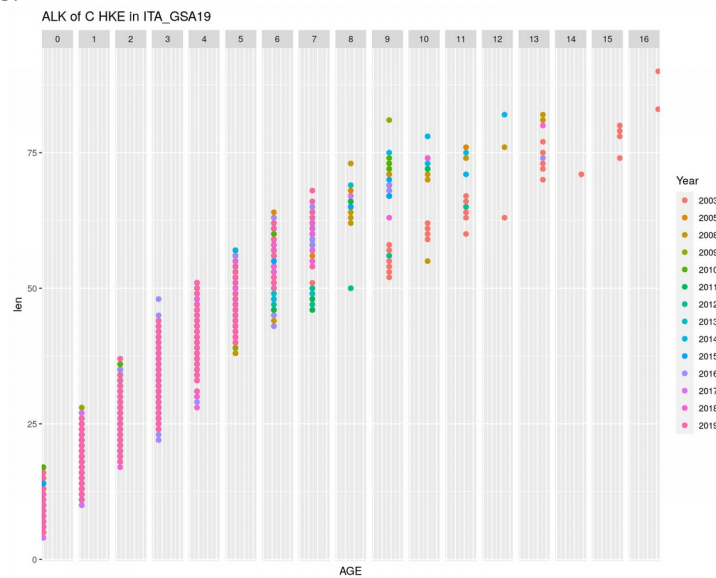
Then for all sex available will be saved:

- plot LW curves facet by year
- plot all curves in one panel
- boxplot of LW parameters available
- trend in time of LW parameters values



Age Length Key (ALK)

Available age keys by sex will be plotted by year (looking for trends) and using a boxplot to find possible outlier values.



Sum Of Product (SOP) landings and discards

Please check the csv files:

SOP_land_needed and SOP_disc_needed.

If there are some issues will be listed as below otherwise the file will be empty.

COUNTRY	YEAR	GEAR	MESH_SIZE_RANGE	FISHERY	AREA	SPECIES	LANDINGS	sumprod	SOPI	ck_nbl	ck_wtl
ITA	2004	GNS	NA	DEF		9 HKE	249.53592	148.125952	1.68461985648538	455.485	3.247
ITA	2005	GNS	NA	DEF		9 HKE	521.70634	226.14228	2.30698275439692	705.775	3.433
ITA	2006	GNS	NA	DEF		9 HKE	592.56606	0.083617	7086.66969635361	1679.964	0.003
ITA	2006	OTB	50D100	DEMSP		9 HKE	327.65157	0.011976	27359.0155310621	5536.356	0.009
ITA	2006	OTB	50D100	MDD		9 HKE	852.31223	0.43797	1946.05162454049	10669.622	0.004
ITA	2007	OTB	50D100	DEMSP		9 HKE	400.42122	0.01743	22973.1049913941	5485.715	0.011
ITA	2008	GNS	NA	DEF		9 HKE	345.23384	0.158434	2179.03884267266	645.642	0.014
ITA	2008	OTB	50D100	DEMSP		9 HKE	789.89715	0.021533	36683.0980355733	12929.719	0.003
ITA	2008	OTB	50D100	MDD		9 HKE	120.75787	0.058579	2061.45325116509	1029.145	0.004
ITA	2009	GNS	NA	DEF		9 HKE	401.2565	0.180414	2224.08737681111	794.55	0.013
ITA	2009	OTB	50D100	DEMSP		9 HKE	762.06229	0.009597	79406.3030113577	9567.071	0.009
ITA	2009	OTB	50D100	MDD		9 HKE	91.17276	0.012869	7084.68101639599	919.432	0.006
ITA	2010	GNS	NA	DEF		9 HKE	576.25583	0.228886	2517.65433447218	1299.405	0.021
ITA	2010	GTR	NA	DEMSP		9 HKE	56.71	0.01161	4884.58225667528	226.336	0.004
ITA	2010	OTB	50D100	DEMSP		9 HKE	617.81163	0.051708	11948.0859828266	7133.561	0.014
ITA	2010	OTB	50D100	MDD		9 HKE	213.11642	0.163472	1303.68760399334	843.978	0.01
ITA	2011	GNS	NA	DEF		9 HKE	502.07963	0.20264	2477.69260757995	1117.528	0.018
ITA	2011	OTB	50D100	DEMSP		9 HKE	696.75993	0.051757	13462.1390343335	7014.332	0.006
ITA	2011	OTB	50D100	MDD		9 HKE	94.626	0.011053	8561.11462951235	579.909	0.003
ITA	2012	GNS	NA	DEF		9 HKE	309.3329	0.141211	2190.57226420038	690.95	0.013
ITA	2012	OTB	50D100	DEMSP		9 HKE	564.95016	0.038561	14650.8171468582	5764.402	0.006
ITA	2012	OTB	50D100	MDD		9 HKE	85.61859	0.017981	4761.61448195317	509.24	0.009
ITA	2013	GNS	NA	DEF		9 HKE	199.21212	0.032476	6134.13351397955	729.394	0.012
ITA	2013	OTB	50D100	DEMSP		9 HKE	963.72521	0.033737	28565.8241693097	10279.291	0.006
ITA	2013	OTB	50D100	DWS		9 HKE	5.35848	0.000796	6731.75879396985	20.769	0.003
ITA	2013	OTB	50D100	MDD		9 HKE	75.2187	0.025009	3007.66524051342	350.16	0.006
ITA	2014	GNS	NA	DEF		9 HKE	177.73147	0.043191	4115.0116922507	490.731	0.014
ITA	2014	OTB	50D100	DEMSP		9 HKE	958.76708	0.447294	2143.4829888172	7415.247	0.007
ITA	2015	GNS	NA	DEF		9 HKE	158.3616	0.066351	2386.72514355473	337.619	0.013
ITA	2015	OTB	50D100	DEMSP		9 HKE	809.016	0.05522	14650.7787033683	7293.054	0.013
ITA	2015	OTB	50D100	MDD		9 HKE	42.86029	0.011532	3716.63978494624	192.511	0.006

The selected rows are based on the threshold set at the beginning of the script

Checking VBGF units

Please have a look in file check_vbunits.csv. If empty no issues have been spotted.

If issues in the VB_units have been spotted, records row number which need to be checked in check_vbunits.csv file (and likely to be fixed in the original GP file) will be printed (If a sequence of NA per Country, Area, Start_year and Species will be visualized it mean that no data were available). If no records are printed the VB units reported should be coherent with the usual VB units used for the analyzed species (cm for Teleosteans and Cephalopods and mm for Crustaceans)

Checking LW parameters units along the years

`table(gp$L_W_UNITS, gp$START_YEAR)`

	2002	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
cm	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

"Please have a look if LW units are coherent with the species and the DCF format. Output AS table of extent 0 x 0 it means no information available"

IV) Filling gaps in Length Frequency Distributions

Data need

DCF format landings/discards by length (see in folder data the file “**landings.csv**” or “**discards.csv**”)

General structure

A folder having two subfolders “**script**” and “**data**” have to be set.

The “**script**” subfolder must be set as working directory folder.

The plots produced will be stored in a new folder (create by the code) called “**output/COMMERCIAL**”

Output/COMMERCIAL folders will have subfolders based on the selected stock:

output/COMMERCIAL/country/gsa/species/Length_frequencies/Landings

for landings and

output/COMMERCIAL/country/gsa/species/Length_frequencies/Discards

for discards data

Main settings

threshold <- 1 # threshold to select which metiers should be included in the reconstruction. For all the metiers will be counted the number of years in which both length frequency distribution (LFD) and yield in weight are available and for how many years at least yield in weight is available. The ratio between these two numbers will be compared to the threshold value. Only if the ratio value will be higher than the threshold the metier will be included in the reconstruction. Higher value (for example 50 or higher) reduce number of metiers take in account in the analysis.

step <- 25 # to be used in plotting LFD (25 should be ok for Teleosteans, 5 for Cephalopods and Crustaceans)

stats <- "median" # you can set "mean" or "median". This stats will be used in creating the "common" LFD by metier which will be used to fill the gap. At the end of the process a SOP correction (actually a rising computation) will be applied to all the reconstructed LFDs.

The a and b parameters to be used in computing weight by length class will be estimated from the available stock data stored in the gp.csv file. If a parameter will not available a warning message will be displayed at the end of the code. In that case enduser has to provided manually a set of a and b values.

Based on the “stats” a median or a mean of the available values will be run to obtained the final set.

The general approach is to work at metier level swapping to upper level (gear) only when no information at metier level are available. Basically if there are enough information on LFD by certain metier the common LFD will be estimated otherwise will be used in starting the process in filling gaps an LFD derived from the corresponding gear distributions.

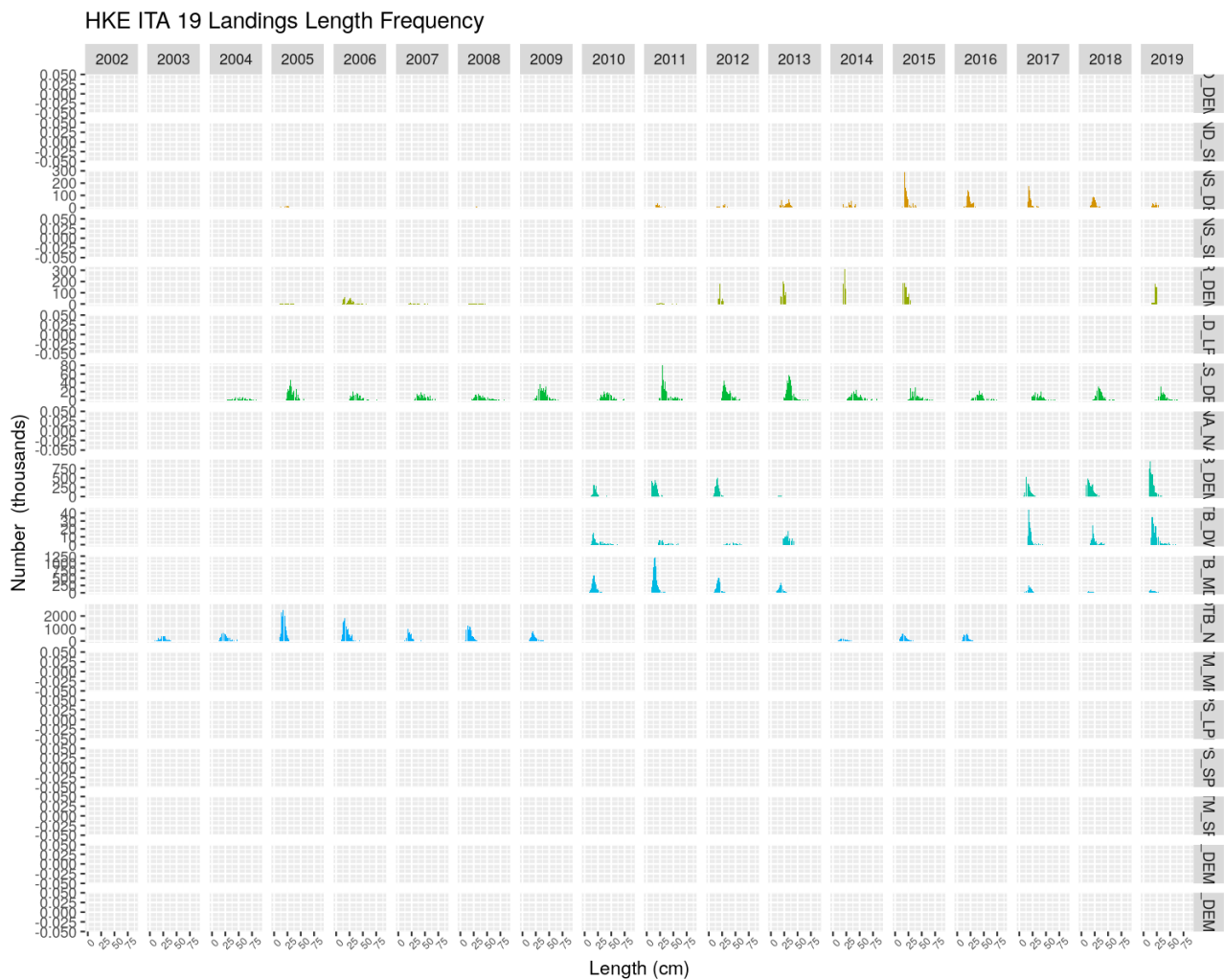
In the discards workflow the missing years will be reconstructed starting to estimating a discards value in weight. This estimation will be run according to a median value for the corresponding metier level (if not possible a ratio estimated on gear will be used). The ratio will be applied to the

landings of the missing years to obtain the discards value in weight. At this point the procedure will follow the same of the landings one.

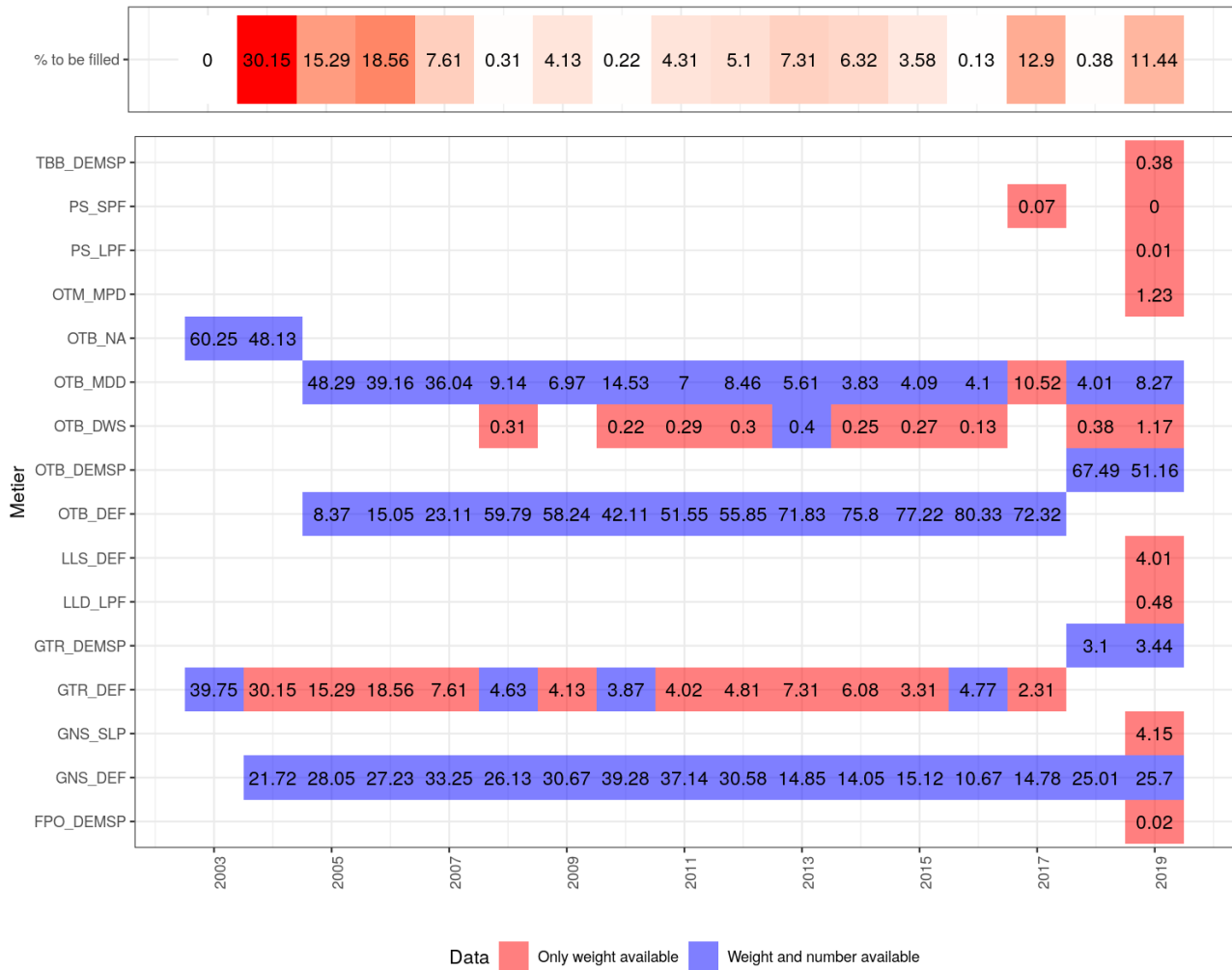
Below are showed the main plots produces by the scripts.

a) Filling gaps in landings data by metier

1) Original length frequency distributions

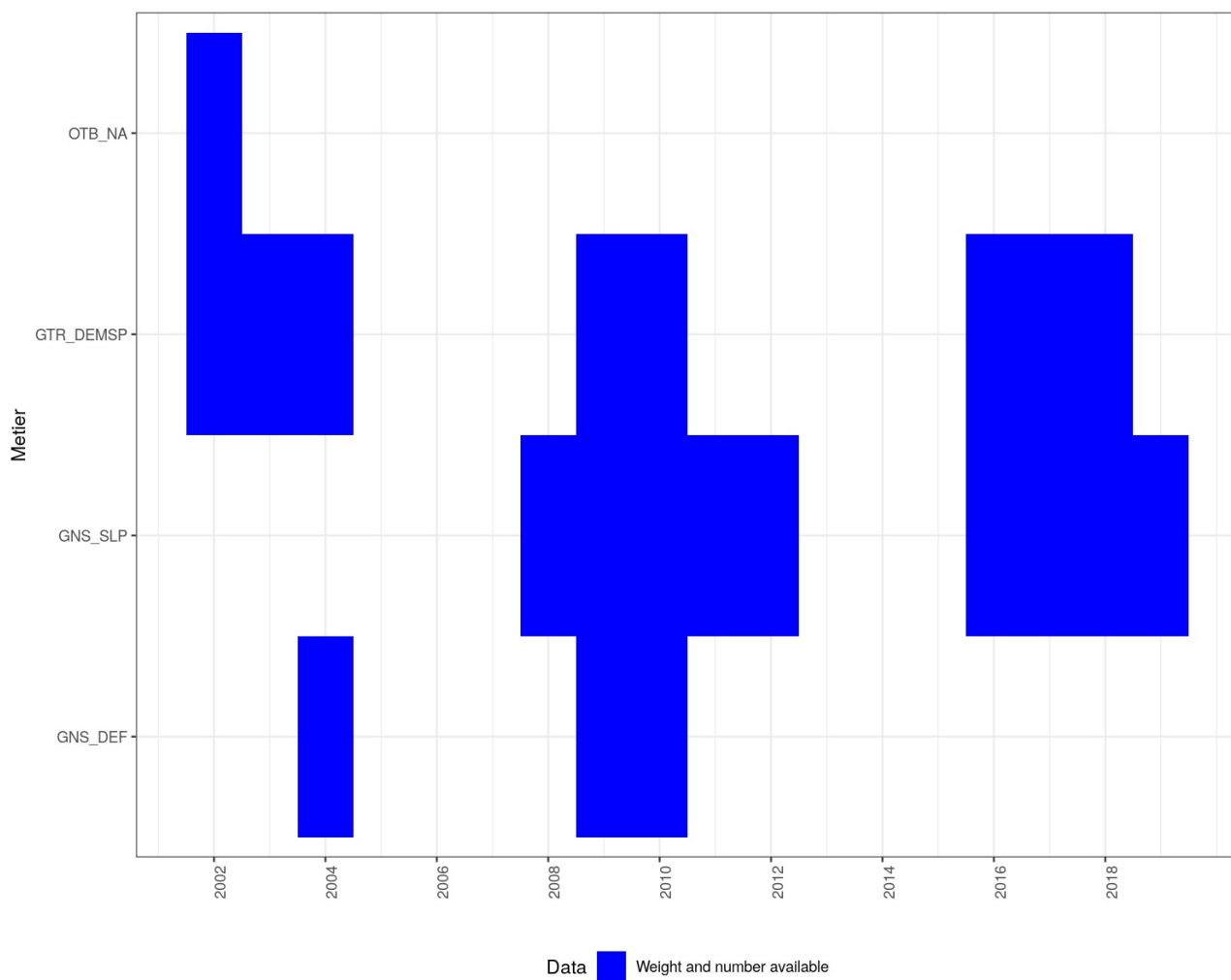


2) Available metier and metier to be reconstructed

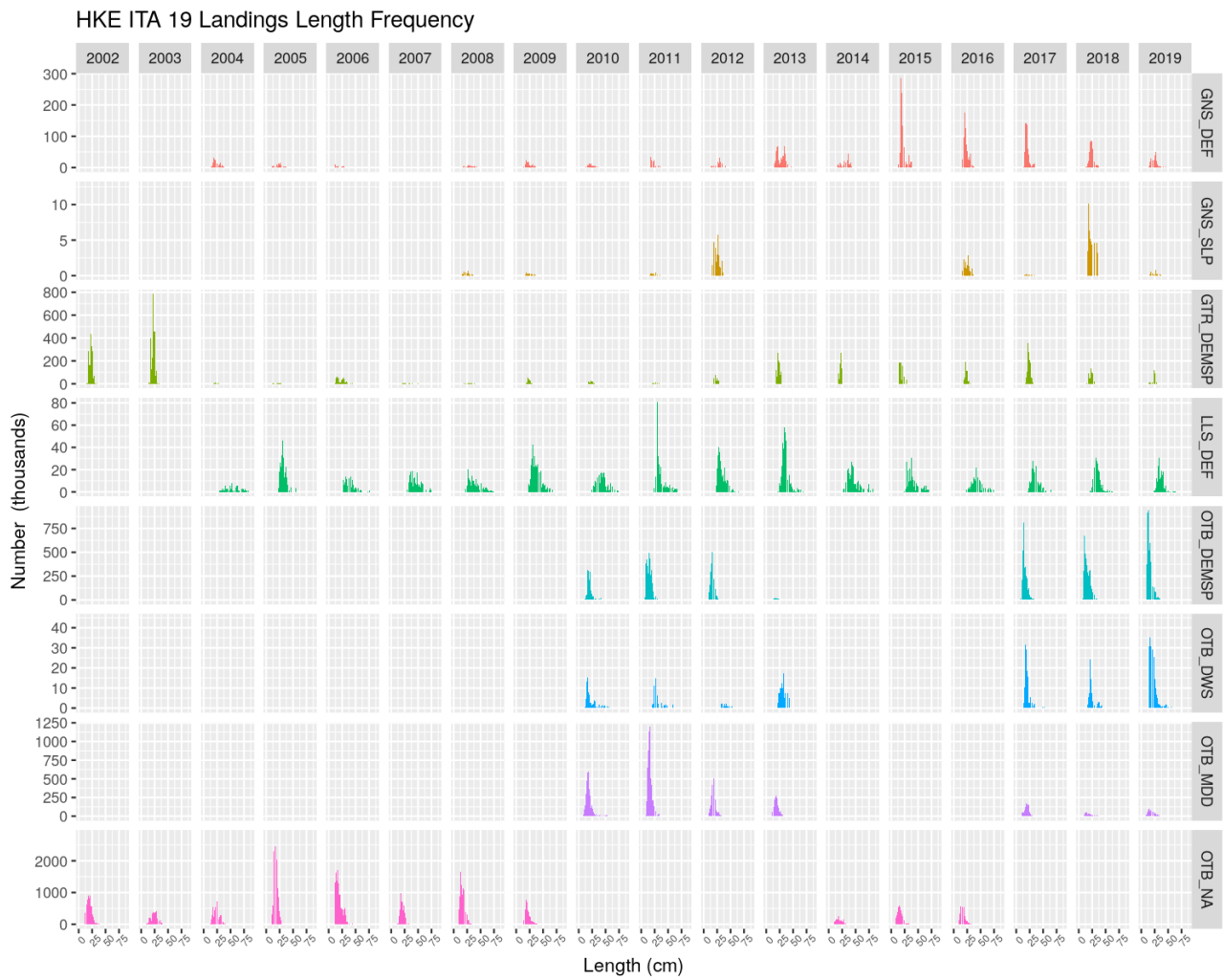


In the upper panel (single row) total percentage of the weight to be reconstructed over total landings by year will be showed (basically it is the sum of the values in the red tiles).

3) Metier reconstructed

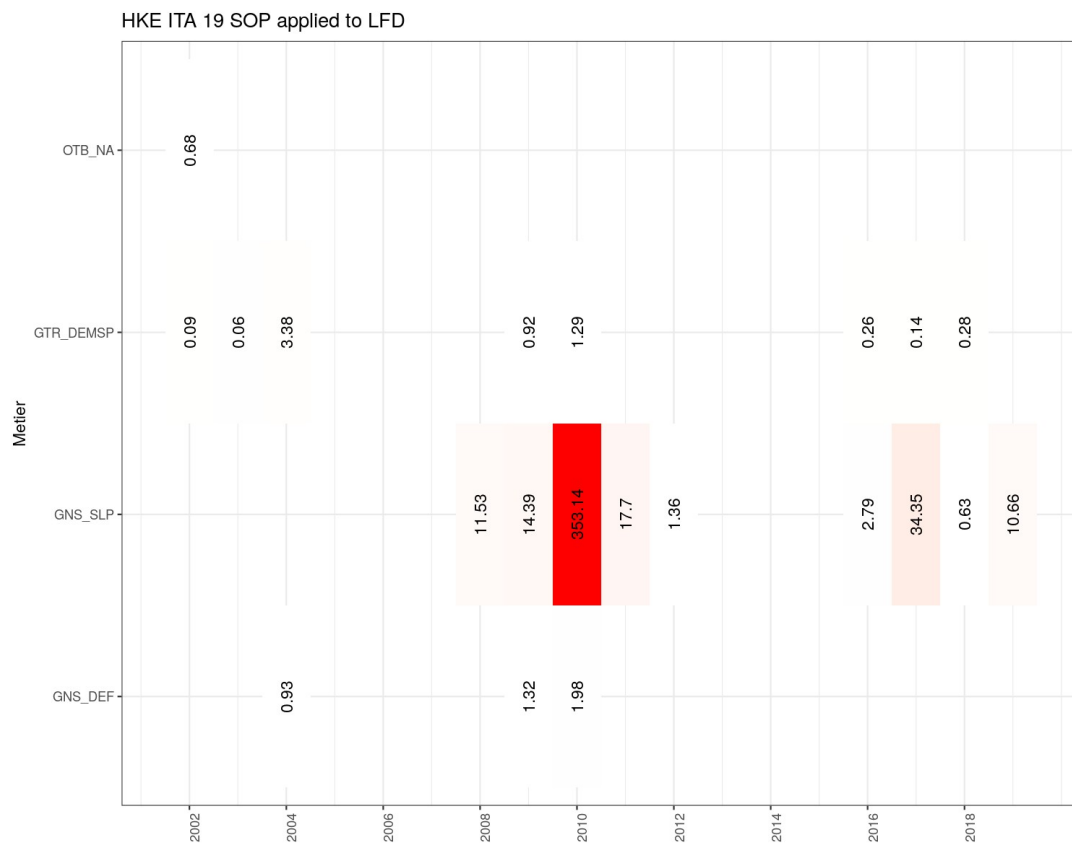


4) Final length frequency distributions



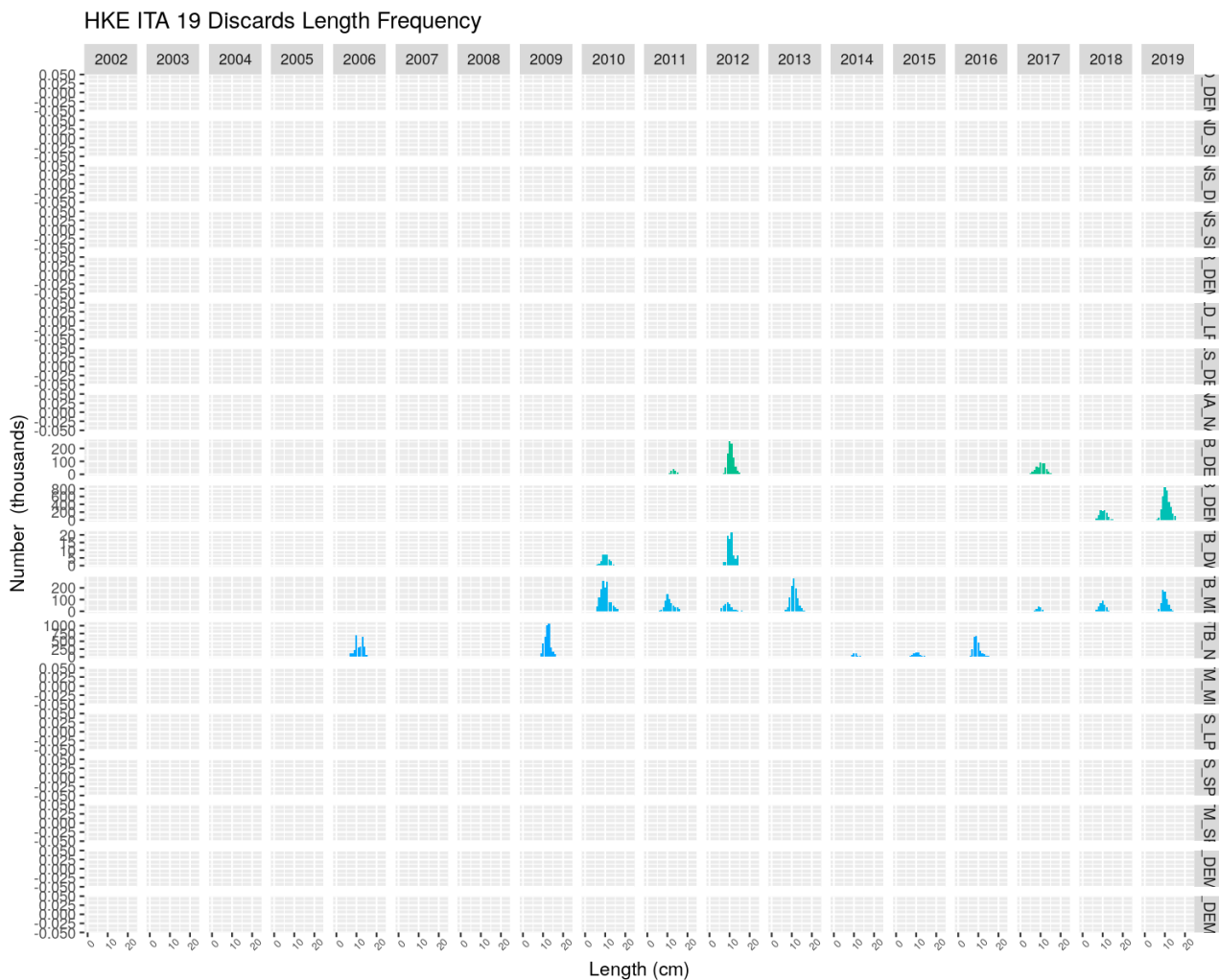
5) Rising factor (aka SOP_ratio) applied in reconstructing missing LFDs

year	metier	total_weight	gear	fishery	yield	SOP_ratio	tot
2002	GTR_DEMSP	23.2000718020684	GTR	DEMSP	262.66353	0.088326201212892	262.66353
2002	OTB_NA	464.820074897213	OTB	NA	687.74508	0.675860996195295	687.74508
2003	GTR_DEMSP	23.2000718020684	GTR	DEMSP	366.6691	0.063272503197211	366.6691
2004	GNS_DEF	32.7945553903196	GNS	DEF	35.11321	0.933966316104954	35.11321
2004	GTR_DEMSP	23.2000718020684	GTR	DEMSP	6.86908	3.37746420220298	6.86908
2008	GNS_SLP	7.26059415152911	GNS	SLP	0.62997	11.5253014453531	0.62997
2009	GNS_DEF	32.7945553903196	GNS	DEF	24.86607	1.31884754568453	24.86607
2009	GNS_SLP	7.26059415152911	GNS	SLP	0.50466	14.3871005261545	0.50466
2009	GTR_DEMSP	23.2000718020684	GTR	DEMSP	25.344	0.915406873503331	25.344
2010	GNS_DEF	32.7945553903196	GNS	DEF	16.59912	1.97568036078537	16.59912
2010	GNS_SLP	7.26059415152911	GNS	SLP	0.02056	353.141738887603	0.02056
2010	GTR_DEMSP	23.2000718020684	GTR	DEMSP	17.94041	1.29317400227021	17.94041
2011	GNS_SLP	7.26059415152911	GNS	SLP	0.41017	17.7014266073314	0.41017
2012	GNS_SLP	7.26059415152911	GNS	SLP	5.32284	1.36404516226847	5.32284
2016	GNS_SLP	7.26059415152911	GNS	SLP	2.60318	2.78912489782847	2.60318
2016	GTR_DEMSP	23.2000718020684	GTR	DEMSP	90.5721	0.256150313419568	90.5721
2017	GNS_SLP	7.26059415152911	GNS	SLP	0.21139	34.3469140050575	0.21139
2017	GTR_DEMSP	23.2000718020684	GTR	DEMSP	167.6923	0.13834905837697	167.6923
2018	GNS_SLP	7.26059415152911	GNS	SLP	11.56713	0.627691929763832	11.56713
2018	GTR_DEMSP	23.2000718020684	GTR	DEMSP	81.493	0.284687909416372	81.493
2019	GNS_SLP	7.26059415152911	GNS	SLP	0.68128	10.6572835714084	0.68128

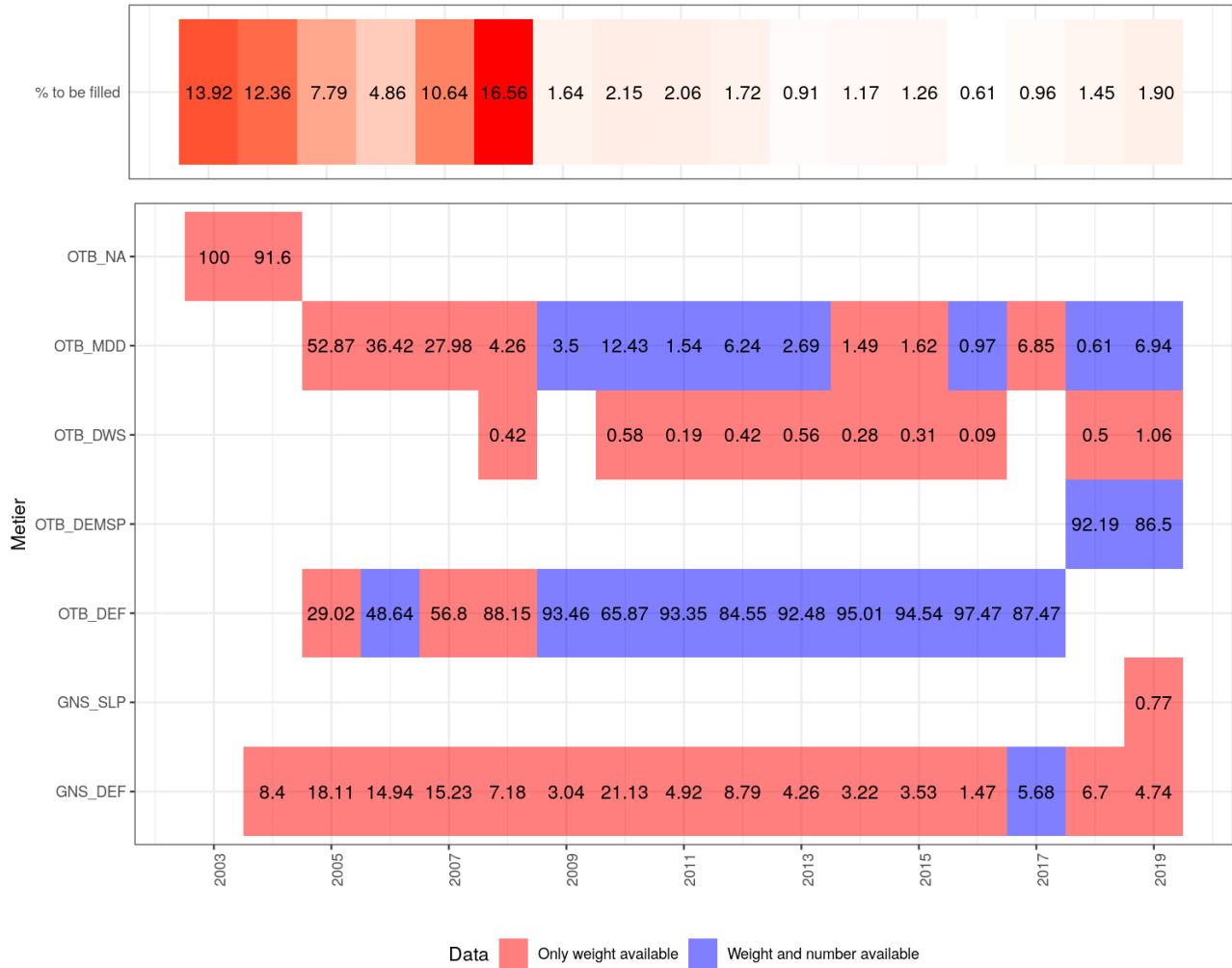


b) Filling gaps in discards data by metier

1) Original length frequency distributions

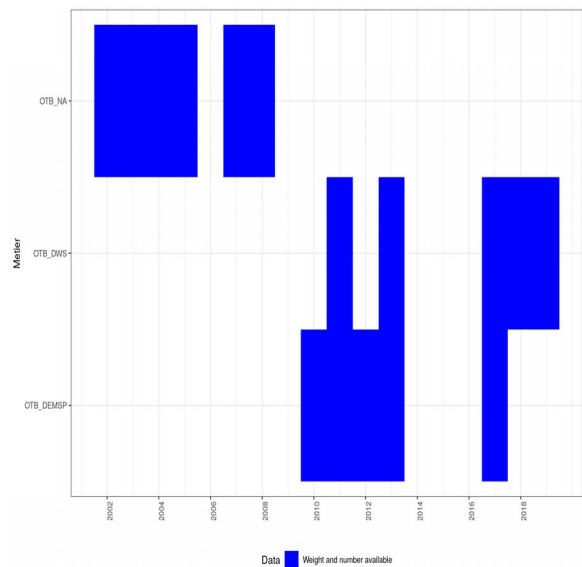


2) Available metier and metier to be reconstructed

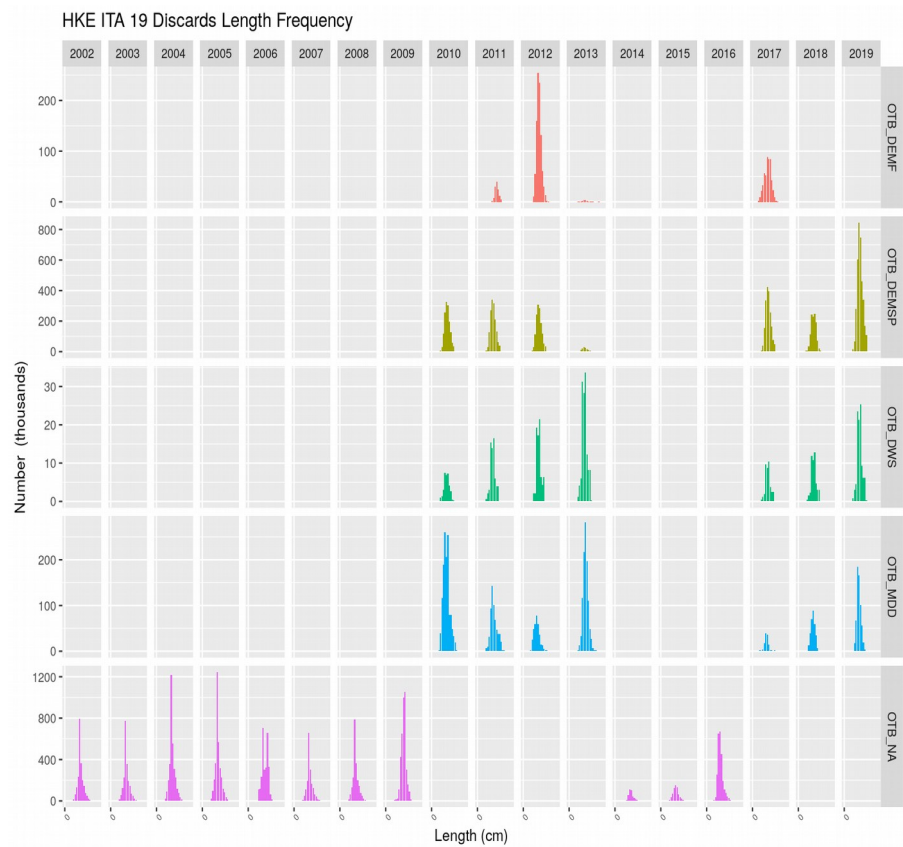


In the upper panel (single row) total percentage of the discards weight to be reconstructed over total catch by year will be showed.

3) Metier reconstructed

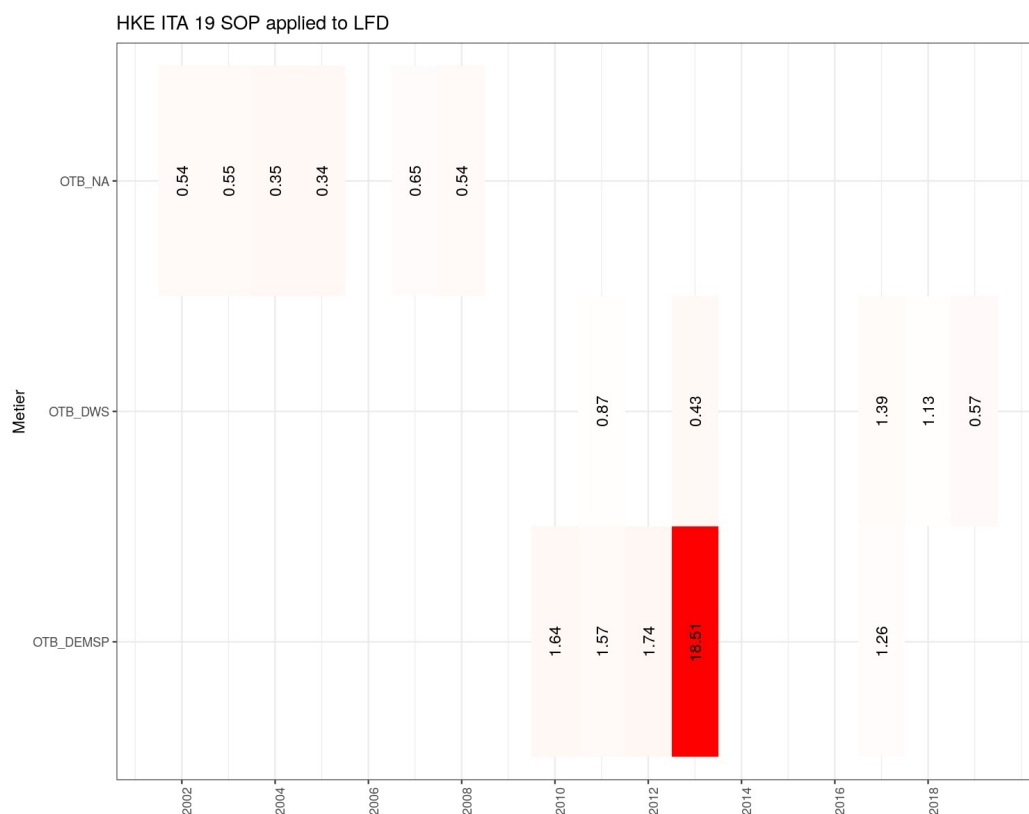


4) Final length frequency distributions



5) Rising factor (aka SOP_ratio) applied in reconstructing missing LFDs

year	metier	total_weight	gear	fishery	yield	SOP_ratio	tot
2002	OTB_NA	9.49791427084486	OTB	NA	17.6754883382859	0.537349468884089	17.6754883382859
2003	OTB_NA	9.49791427084486	OTB	NA	17.1795785655981	0.552860725574743	17.1795785655981
2004	OTB_NA	9.49791427084486	OTB	NA	27.0659548286649	0.350917391644571	27.0659548286649
2005	OTB_NA	9.49791427084486	OTB	NA	27.712328224025	0.342732454453636	27.712328224025
2007	OTB_NA	9.49791427084486	OTB	NA	14.6879221309113	0.646647918350282	14.6879221309113
2008	OTB_NA	9.49791427084486	OTB	NA	17.5273539645675	0.54189093744814	17.5273539645675
2010	OTB_DEMSP	20.1412077769111	OTB	DEMSP	12.2575994835379	1.64316086554802	12.2575994835379
2011	OTB_DEMSP	20.1412077769111	OTB	DEMSP	12.8691934007233	1.56507149669373	12.8691934007233
2011	OTB_DWS	0.452162661401024	OTB	DWS	0.520217680694153	0.86917972645159	0.520217680694153
2012	OTB_DEMSP	20.1412077769111	OTB	DEMSP	11.5703410630843	1.74076180357143	11.5703410630843
2013	OTB_DEMSP	20.1412077769111	OTB	DEMSP	1.08832557522764	18.5066015495393	1.08832557522764
2013	OTB_DWS	0.452162661401024	OTB	DWS	1.05866351536325	0.42710705983466	1.05866351536325
2017	OTB_DEMSP	20.1412077769111	OTB	DEMSP	15.9238115510914	1.2648484134774	15.9238115510914
2017	OTB_DWS	0.452162661401024	OTB	DWS	0.325861710575588	1.38759064574461	0.325861710575588
2018	OTB_DWS	0.452162661401024	OTB	DWS	0.40064062680038	1.12859912638444	0.40064062680038
2019	OTB_DWS	0.452162661401024	OTB	DWS	0.797556015936488	0.56693530280766	0.797556015936488



6) Metier mismatch

Please check if there are any mismatch in the metier reported in discards and declared in landings.
Discards in weight is no estimable without a value in landings.

year	gear	fishery	land	discards	Data
2011	OTB	DEMF	NA	1.983722	Metier mismatch
2012	OTB	DEMF	NA	8.532566	Metier mismatch
2013	OTB	DEMF	NA	0.159404	Metier mismatch
2017	OTB	DEMF	NA	4.613141	Metier mismatch

V) Relative weights

Data need

DCF format landings/discards by length (see in folder data the file “**landings.csv**” or “**discards.csv**”)

General structure

A folder having two subfolders “**script**” and “**data**” have to be set.

The “**script**” subfolder must be set as working directory folder.

The plots produced will be stored in a new folder (create by the code) called “**output/COMMERCIAL**”

Output/COMMERCIAL folders will have subfolders based on the selected stock:

output/COMMERCIAL/country/gsa/species/Landing

for landings and

output/COMMERCIAL/country/gsa/species/Discard

for discards data

Stock selection

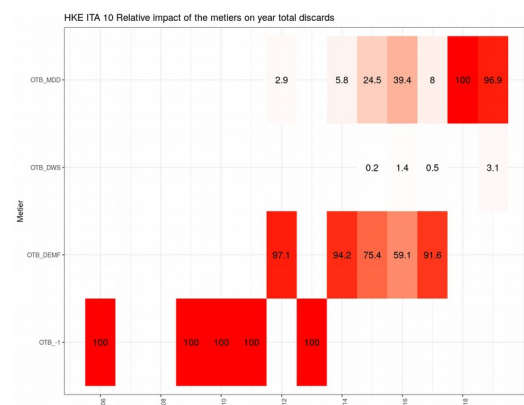
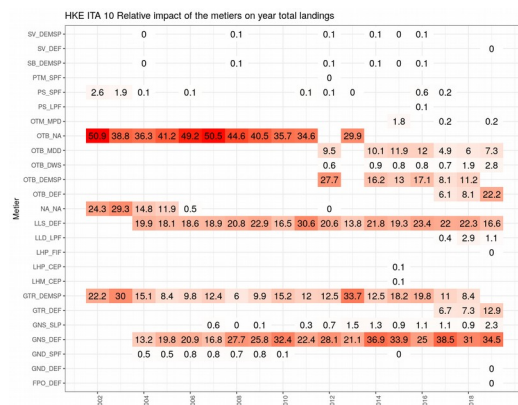
The code works at stock level as we need to declare for which country, gsa and species we want perform the analysis.

Choose Member state and area to check

MS <- "ITA"

GSA <- "11"

SP <- "ARA"



Acknowledgment

I want to thank all the experts involved in the STECF EWG2101 for all suggestions and help in improving the R scripts suite.