Call MARE/2020/08

Agreement reference: SI2.839444
European Maritime and Fisheries Fund (EMFF)

Development of the Regional Database for the Mediterranean & Black Seas

Work-package 4 - Deliverable 4.3.1

R package containing a priori and a posteriori quality checks with pertinent documentation

I. Bitetto, G. Tserpes

Partners involved:

COISPA, HCMR, CIBM, CNR

Table of Contents

Acronyms

CP	Contracting Parties
CRAN	Comprehensive R Archive Network
DCRF	Data Collection Reference Framework
EWG	Expert Working Group
FDI	Fishery Dependent Data
GFCM	General Fisheries Commission for the Mediterranean
MED&BS	Mediterranean and Black Sea
STECF	Scientific, Technical and Economic Committee for Fisheries

Executive summary

In the last years, the data quality issue gained increasing importance in the EU Data Collection Framework. Some of the main objectives in fishery data management is that data should be Findable, Accessible, Interoperable, and Reusable.

For this reason the European Commission devoted several resources aimed at improving the data quality, through the development of free and transparent tools (STREAM project, MARE/2016/22 – SI2.770115) and dedicated working groups to data quality (STECF EWG 22-03, EWG 21-02).

On the other hand, also GFCM made progress in implementing fisheries data quality indicators within the DCRF, namely timeliness, completeness, conformity, stability and consistency[1] (16t Liaison Meeting RCG, held in 2019 in Brussels).

Task 4.3 - Developing data validation and quality checking tools - built on the experience gained in STREAM and STECF EWG concerning data quality procedures to be applied both on detailed and aggregated data; the existing quality checks have been enhanced and extended to the FDI and GFCM data calls, in order to provide an all-round tool freely downloadable and easy to use for a basic R user.

Concerning the survey data, the RoME initiative has been taken into account to improve the quality of this type of data. RoME was presented for the first time in the MEDITS Coordination meeting held in Nantes (March 2011) and represents a common tool to perform the data checks with a standardized procedure within MEDITS (Spedicato et., 2019, MEDITS Handbook, 2017) coordination.

The availability of such tools is expected to dramatically reduce the risk of data failure of the Member States in the data calls data submission.

This task provided two complete, open-access (available on GitHub repository) and well documented tools for fishery dependent and fishery independent data to dramatically reduce the risk of data failure of the MS and CPs in the data submission to the four main data calls present in Mediterranean and Black Sea: MED&BS, FDI, GFCM DCRF and RCG.

In the development of this task contacts and communications with RCG and end users (STECF and GFCM) has been implemented for identifying specific needs. This communication is expected to be maintained in the future, to continue the testing of the packages, as well as their adaptation of them to specific future data needs of end-users.

In this perspective the storing on the GitHub repository and the development in R language will facilitate this code adaptation and improvement, allowing the collaboration among the experts involved in the data calls.

[1] *https://datacollection.jrc.ec.europa.eu/documents/10213/1239605/2019-12_16th_Liaison_Meeting.pdf/c59331f6-3047-4f25-a0b0-89945475a174?version=1.2*

1. Introduction

In the last years, the data quality issue gained increasing importance in the EU Data Collection Framework. Some of the main objectives in fishery data management in are laid out in the FAIR data principles, that data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).

For this reason, the European Commission devoted several resources aimed at improving the data quality, through the development of free and transparent tools (STREAM project, MARE/2016/22 – SI2.770115) and dedicated working groups to data quality (STECF EWG 22-03, EWG 21-02).

On the other hand, also GFCM made progress in implementing fisheries data quality indicators within the DCRF, namely timeliness, completeness, conformity, stability and consistency[2] (16t Liaison Meeting RCG, held in 2019 in Brussels).

Task 4.3 - Developing data validation and quality checking tools - built on the experience gained in STREAM and STECF EWG concerning data quality procedures to be applied both on detailed and aggregated data; the existing quality checks have been enhanced and extended to the FDI and GFCM data calls, in order to provide an all-round tool freely downloadable and easy to use for a basic R user.

Concerning survey data, the RoME initiative has been considered to improve the quality of this type of data. RoME was presented for the first time in the MEDITS Coordination meeting held in Nantes (March 2011) and represents a common tool to perform the data checks with a standardized procedure within MEDITS (Spedicato et., 2019, MEDITS Handbook, 2017) coordination.

The availability of such tools is expected to dramatically reduce the risk of data failure of the Member States in the data calls data submission.

2. Working method

RDBqc is an a hoc R data validation and quality check tool based on the concept of a 2-steps process (Figure 2-1) to verify the consistency of the biological data and the aggregated data:

- *a priori* quality checks (QC), to detect possible inconsistency and inaccuracies already present in the detailed data;
- *a posteriori* QC, designed to verify the temporal and spatial coverage, as well as that the data consistency is maintained in the aggregated dataset.

Both the *a priori* and the *a posteriori* QC were originally implemented through an Rmd script, in STREAM project, allowing to produce an automatic report at the end of the procedure, indicating the outcomes of each quality check. However, a first innovation of this task has been to disentangle the actual QC procedures from the automatic production of report, in order to deliver a comprehensive R library of data validation and quality check functions.
In the development of this task contacts and communications with RCG and end users (STECF and GFCM) has been implemented for identifying specific needs.

The last version of the package can be found on a publicly accessible GitHub repository: *https://github.com/COISPA/RDBqc/tree/main/RDBqc*.

[2] *https://datacollection.jrc.ec.europa.eu/documents/10213/1239605/2019-12_16th_Liaison_Meeting.pdf/c59331f6-3047-4f25-a0b0-89945475a174?version=1.2*

To support the quality check of the MEDITS survey data, a new updated version (v.0.1.23) of RoME was specifically developed in RDBFIS project to be integrated in RDBFIS environment. The main function RoMEcc was developed, allowing to check all the tables without any interruption and reporting all the errors detected at the end of the analysis. Nevertheless, the new application can be run also as standalone.

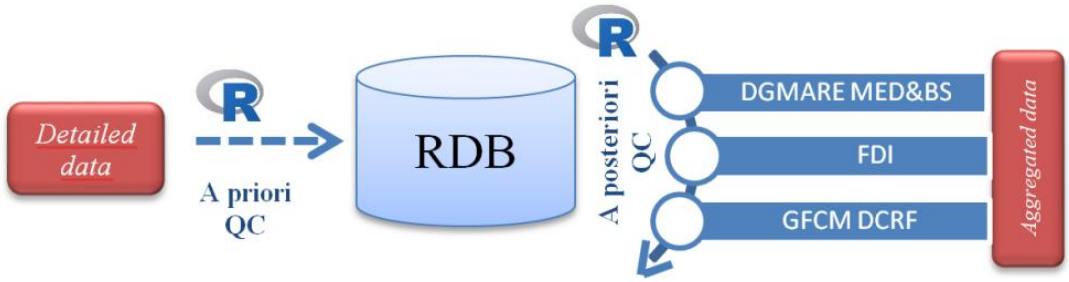

Figure 2-1 General scheme of two steps approach implemented in RDBqc package.

2.1 Data quality checks on RCG CS (sampling) and CL (landing) tables

Respect to STREAM project, additional checks have been implemented to further improve the data quality of detailed data. Specifically, the new functions are aimed at:

- summarizing the number of individual biological data (length, sex, maturity, weight and age) collected by trip;
- summarizing the number of trips/hauls monitored by year by port, metier, sampling method;
- localizing the hauls position or the average trip position by year in a map.

These new features have been thought to answer to the proposals made for the Med & BS RDB by the 59 ICES Steering Committee of the Regional Fisheries Database (SCRDB; ICES, 2020), concerning the automatic completion of Annual Reports tables and the geo-referred presentation of the data and automatized data-quality. The *a priori* quality checks are applied before storing the detailed data in the database (in the uploading phase).

Moreover, some checks havebeen implemented for CL (landing) table of RCG format, specifically:

- 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 metier;
- qualitative checks of time series, through graphical outputs.

In Table 2.1-1 are listed all quality checks implemented in RDBqc package as single functions; in the table is also indicated the datacall's table on which the check is applied and a short description of the check.

A detailed description about the functions can be found in **ANNEX XIV RDBqc_0.0.14**.

Table 2.1-1 – List of the quality checks functions implemented for RCG data call (CS and CL tables).

Qualitycheck	Datacall	Table	Checkdescription
check_CL	RCG MED & BS	CL	Spatial and temporal coverage of landing and landing values in CL tables
check_AL	RCG MED & BS	CS	Consistency of length-age data: plot, summary table and list of errors (outliers).
check_LFD	RCG MED &	CS	Consistency of length data: plot and error

	BS		(outliers)
check_LFD_comm_cat	RCG MED & BS	CS	Consistency of length and commercial category: plot and a summary table with ranges by year and commercial category
check_loc	RCG MED & BS	CS	Visual check of haul and trip position
check_lw	RCG MED & BS	CS	Consistency of length-weight data: visual check and errors (outliers).
check_mat	RCG MED & BS	CS	Consistency between length and maturity stage: plot of the maturity stages by length class
check_mat_ogive	RCG MED & BS	CS	Consistency of maturity stages and length, through the estimation of maturity ogive by sex
summarize_ind_meas	RCG MED & BS	CS	Summary on the number of individuals by trip for which biological data have been collected (length, sex, maturity, weight and age)
summarize_trips	RCG MED & BS	CS	Summary on the number of trips/hauls monitored by year by port, metier, sampling method

2.2 Data quality checks on MED&BS, FDI, GFCM DCRF data calls

For the *a posteriori* QC, the scripts developed in STREAM have been utilized as a basis for implementing new functions. In particular, the script already developed, allowed to:

i) verify the temporal and spatial coverage of all the tables requested under DGMARE Med&BS datacall;
ii) verify the sum of products in the Catch table;

the consistency of the biological information along the years by species and area, will be translated in R functions.

In Table 2.2-1 is reported the extended list of all the quality check functions implemented in the RDBqc package for MED&BS data call.

A detailed description about the functions can be found in **ANNEX XIV RDBqc_0.0.14**.

Table 2.2-1 – List of the quality checks functions implemented for MED & BS data call.

Qualitycheck	Datacall	Table	Checkdescription
MEDBS_Catch_coverage	MED & BS	Catch	Summary tables and plot of landing and discards by year, Country, quarter, GSA, LOA, gear, mesh size range and fishery.
MEDBS_Discard_coverage	MED & BS	Discard	Summary table and plot of discards by year, Country, quarter, GSA, LOA, gear, mesh size range and fishery.
MEDBS_check_duplicates	MED & BS	Landing/ Discard/ Catch	Check duplicates in landing at length table

MEDBS_comp_disc_YQ	MED & BS	Discard	Comparison of discards aggregated by quarters and by year
MEDBS_comp_disc_YQ_fishery	MED & BS	Discard	Comparison of discards aggregated by quarters and by year and fishery
MEDBS_comp_land_Q_VL	MED & BS	Landing	Comparison of landings aggregated by quarters accounting for the presence of vessel length
MEDBS_comp_land_Q_VL_fishery	MED & BS	Landing	Comparison of landings aggregated by quarters and fishery accounting for the presence of vessel length
MEDBS_comp_land_YQ	MED & BS	Landing	Comparison of landings aggregated by quarters and by year
MEDBS_comp_land_YQ_fishery	MED & BS	Landing	Comparison of landings aggregated by quarters and by year and fishery
MEDBS_disc_mean_weight	MED & BS	Discard	Consistency of mean weight discard: plot of the discards weight by year, gear and fishery
MEDBS_ks	MED & BS	Landing/ Discard	Kolmogorov-Smirnov test on cumulative landing and discard function at length
MEDBS_land_mean_weight	MED & BS	Landing	Consistency of mean weight: plot and data frame of mean weight by year,gear and fishery
MEDBS_length_ind	MED & BS	Landing/ Discard	Consistency of length data: Main length size indicators
MEDBS_lengthclass_0	MED & BS	Landing/ Discard	Detection of the records with null individuals in landings and discards
MEDBS_plot_disc_vol	MED & BS	Discard	Consistency of time series of discard: Plot of total discards by gear and fishery.
MEDBS_plot_discard_ts	MED & BS	Discard	Consistency of time series of discard: plot by year or by quarter.
MEDBS_plot_land_vol	MED & BS	Landing	Consistency of time series of landing: Plot of total discards by gear and fishery.
MEDBS_plot_landing_ts	MED & BS	Landing	Consistency of time series of landing: plot by year or by quarter.
MEDBS_weight_0	MED & BS	Landing/ Discard	Consistency of weight in landing and discard at length tables: weight 0 in landings and discards
MEDBS_weight_minus1	MED & BS	Landing/ Discard	Consistency of weight in landing and discard at length tables: weight -1 in landings and discards
MEDBS_yr_missing_length	MED & BS	Landing/ Discard	Detection of records with years with missing length distributions
MEDBS_GP_check	MED & BS	GP	Consistency of growth parameters in GP table
MEDBS_Landing_coverage	MED & BS	Landing	Summary table and plot of landings by year, Country, quarter, GSA, LOA, gear, mesh size range and fishery.

MEDBS_LW_check	MED & BS	GP	Consistency of length-weight relationship parameters in GP table
MEDBS_MA_check	MED & BS	MA	Consistency of maturity ogives at age across the years
MEDBS_ML_check	MED & BS	ML	Consistency of maturity ogives at length across the years
MEDBS_SA_check	MED & BS	SRA	Consistency of sex ratio at age across the years
MEDBS_SL_check	MED & BS	SRL	Consistency of sex ratio at length across the years
MEDBS_ALK	MED & BS	ALK	Plot of Age-Length Keys
MEDBS_SOP	MED & BS	Catch	check of the sum of products

Moreover, additional functions have been specifically implemented on the FDI (Fishery Dependent Information) tables, relevant for Mediterranean and Black Sea (specifically G, H, I and J tables) for RDBFIS project. These new functions are aimed at verifying the consistency of transversal variables (effort and landing) as well as the spatial and temporal coverage. In Table 2.2-2 is reported the extended list of all the quality check functions implemented in the RDBqc package for FDI data call.

Table 2.2-2 – List of the quality checks functions implemented for FDI data call.

Qualitycheck	Datacall	Table	Checkdescription
FDI_check_coord	FDI	H, I	Compatibility of the geographical coordinates with rectangle type
FDI_Check_empty_field_A	FDI	A	Detection of emptyfields
FDI_Check_empty_field_G	FDI	G	Detection of emptyfields
FDI_Check_empty_field_H	FDI	H	Detection of emptyfields
FDI_Check_empty_field_I	FDI	I	Detection of emptyfields
FDI_Check_empty_field_J	FDI	J	Detection of emptyfields
FDI_Check_record_duplicated_A	FDI	A	Detectionduplicatedrecords
FDI_Check_record_duplicated_G	FDI	G	Detectionduplicatedrecords
FDI_Check_record_duplicated_H	FDI	H	Detectionduplicatedrecords
FDI_Check_record_duplicated_I	FDI	I	Detectionduplicatedrecords
FDI_Check_record_duplicated_J	FDI	J	Detectionduplicatedrecords
FDI_checks_spatial_HI	FDI	H, I	check NA values in spatial columns of both table H and I
FDI_cov_tableA	FDI	A	Coverage by year, GSA, MS, species, vessels length, and fishing techniques for Total live weight landed, total value of landings, and total discards.

FDI_cov_tableG	FDI	G	Coverage by year, GSA, MS, vessels length, fishing techniques, and metier for total days at se, total Fishing Days, total kW days at Sea, total GT days at sea, total kW fishing days, totgtfishdays, hours at Sea, kW hours at sea.
FDI_cov_tableJ	FDI	J	Check number of record in FDI J table grouped by year, GSA, MS, vessels length, and fishing techniques for total trips; total kW; total GT; total vessels.
FDI_coverage	FDI	A, G, H, I, J	Coverage by GSA and year
FDI_Cross_checks_AG	FDI	A, G	Cross check between landing and effort in FDI tables A and G
FDI_Cross_Checks_AH	FDI	A, H	Cross check between landings in table A and spatial landings in table H
FDI_Cross_Checks_IG	FDI	I, G	Check consistency between spatial effort in table I and effort in table G
FDI_Cross_Checks_JG	FDI	J,G	Check consistency between amount of vessels in table J capacity and the amount of vessels in table G
FDI_disc_coverage	FDI	A	Coverage of FDI discard data
FDI_fishdays_cov	FDI	G, I	Coverage comparison of totfishdays between FDI tables G and I
FDI_landweight_cov	FDI	A, H	Coverage of weight of landings in FDI table A and H
FDI_price_cov	FDI	A	Check the trend prices in the given table grouped by year, GSA, MS, and species.
FDI_prices_not_null	FDI	A	Detection of cases with total landings > 0 but landings value = 0.
FDI_vessel_length	FDI	J	Check of vessel length in FDI table
FDI_vessel_numbers	FDI	J, G	Check number of vessels in FDI table J and G

Finally, *ad hoc* functions have been implemented for *a posteriori* QC specific for the GFCM DCRF tasks on biological information.

In Table 2.2-3 is reported the extended list of all the quality check functions implemented in the RDBqc package for GFCM DCRF data call.

Table 2.2-3 – List of the quality checks functions implemented GFCM DCRF data call.

Qualitycheck	Datacall	Table	Checkdescription
GFCM_Check_empty_field_II2.R	GFCM DCRF	task II.2	Detection of emptyfields
GFCM_Check_empty_field_III.R	GFCM DCRF	task III	Detection of emptyfields
GFCM_Check_empty_field_VII2.R	GFCM DCRF	task VII.2	Detection of emptyfields
GFCM_Check_empty_field_VII31.R	GFCM DCRF	task VII.3.1	Detection of emptyfields
GFCM_Check_empty_field_VII32.R	GFCM DCRF	task VII.3.2	Detection of emptyfields
GFCM_Check_L50_VII31.R	GFCM DCRF	task VII.3.1	Consistency of L50 values in Task VII.3.1 table
GFCM_Check_lfd_VII2.R	GFCM DCRF	task VII.2	Check the consistency of the length frequency distributions (LFD) reported in the TaskVII.2 table
GFCM_Check_min_max_L50_VII31	GFCM DCRF	task VII.3.1	Check the consistency of L50 reported in the TaskVII.3.1 table with the theoretical values reported in the minmaxLtaskVII31 table
GFCM_Check_min_max_length_VII2	GFCM DCRF	task VII.2	Check the consistency of the lengths reported in the TaskVII.2 table with the theoretical values reported in the minmaxLtaskVII2 table
GFCM_Check_record_duplicated_II2.R	GFCM DCRF	task VII.2	Detectionduplicatedrecords
GFCM_Check_record_duplicated_III.R	GFCM DCRF	task III	Detectionduplicatedrecords
GFCM_Check_record_duplicated_VII2.R	GFCM DCRF	task VII.2	Detectionduplicatedrecords
GFCM_Check_record_duplicated_VII31.R	GFCM DCRF	task VII.3.1	Detectionduplicatedrecords
GFCM_Check_record_duplicated_VII32.R	GFCM DCRF	task VII.3.2	Detectionduplicatedrecords
GFCM_check_species_catfau_VII32.R	GFCM DCRF	task VII.3.2	Check mismatching species/Catfau and Sex per maturity stages for Task VII.3.2 table

GFCM_cov_task2.2	GFCM DCRF	task II.2	Coverage by year, GSA, MS, species, and segment for Total landing and total discards.
GFCM_cov_task3	GFCM DCRF	task III	Check number of individuals in GFCM Task III table
GFCM_relatioship_length_weight_VII2.R	GFCM DCRF	task VII.2	Check the consistency of length-weight relationship in the GFCM Task VII.2 table
GFCM_relatioship_length_weight_VII32.R	GFCM DCRF	task VII.3.2	check the consistency of length-weight relationship in the GFCM Task VII.3.2 table by sex
GFCM_Check_presence_GSA_FS_II2	GFCM DCRF	task II.2	Check of missing combination GSA/Fleet segment per year
GFCM_lmat_TaskVII32	GFCM DCRF	task VII.3.2	plot the lengths at maturity stages by species and sex to easily identify outliers

2.3 Data quality checks on MEDITS survey data

RoME works on the tables in the MEDITS format and uses some tables included in the MEDITS manual. This tool foresees a lot of checks and cross checks based on the TA, TB, TC, TE and TL tables (Figure 2.3-1). A detailed description about the function working on TA, TB and TC can be found in **ANNEX XV RoME_0.1.23** and in the interactive help of the package (Figure 2.3-2).

The new RoME version is compatible with the more recent versions of R; the outputs are text files (logfile) reporting the outcome of each check and plots for qualitative controls. RoME is multiplatform, with 67 standalone check functions, each one with documentation included. Furthermore, the quality of plots and maps has been globally improved. New functions were developed specifically to work on RDBFIS.

The report of RoME is represented by:

- a zip file, containing a logfile, all the graphical and tabular outcomes;
- a data frame where all the outcomes of the quality checks are schematized (it is a return of the main R function).

RoME can be freely downloadable from GitHub: *https://github.com/COISPA/RoME*.

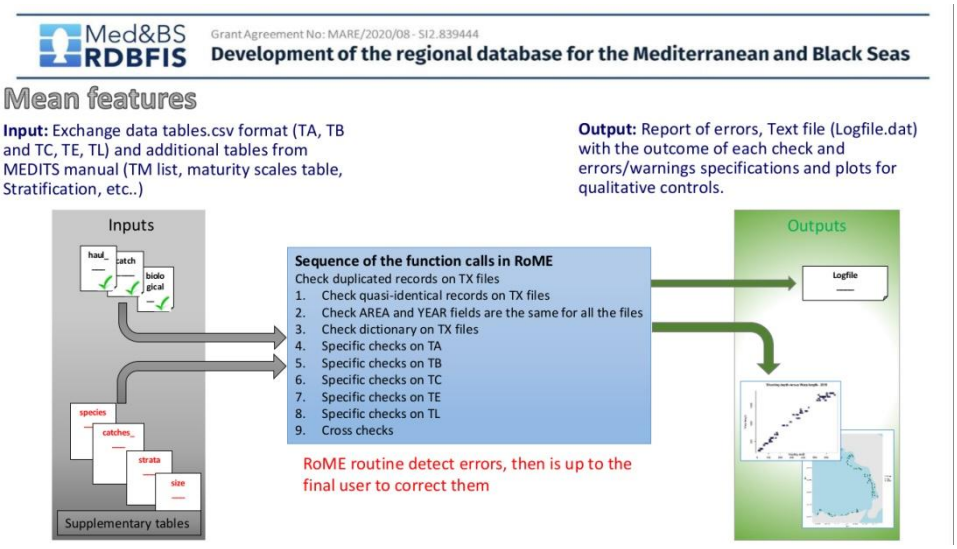

Figure 2.3-1 Scheme of the new RoME version functioning.

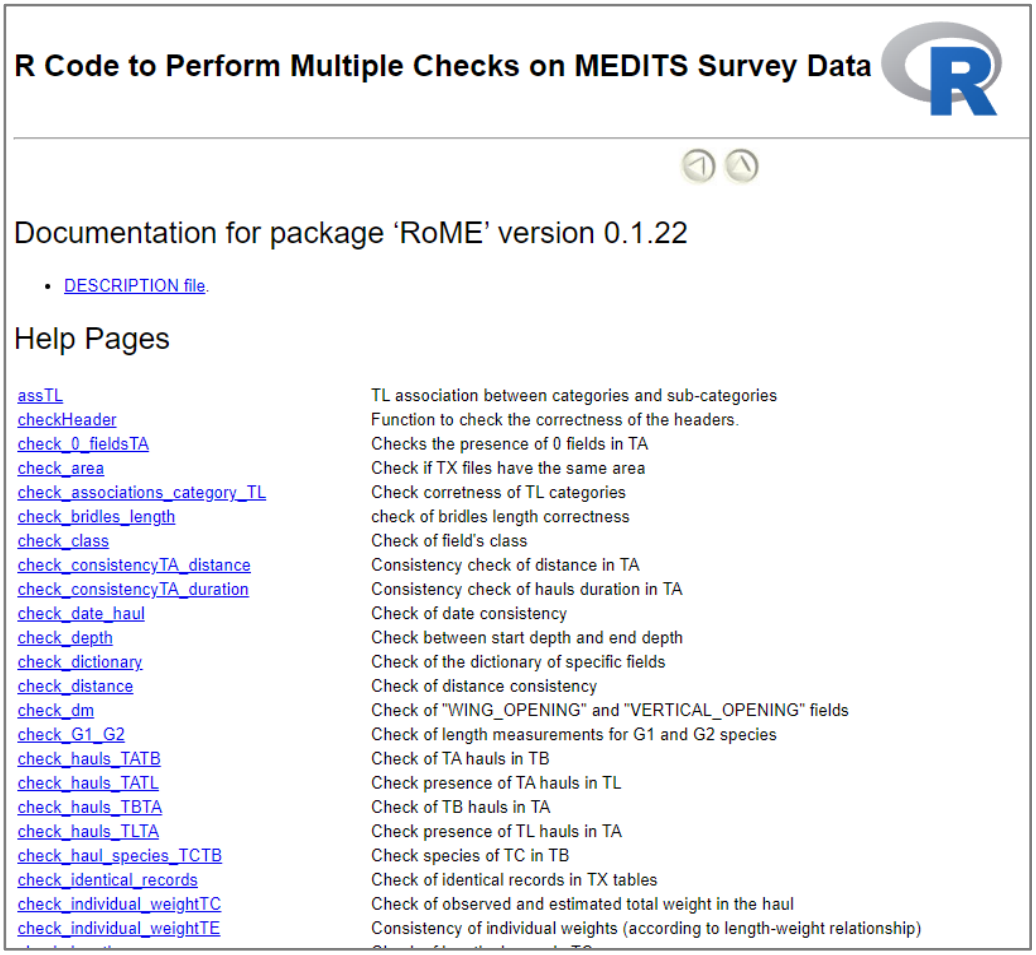

Figure 2.3-2Interactive help in RoME package.

The extended list of the checks implemented in h the new version of RoME follows.

2.3.1 TA table (haul data)

1. Check the presence of 0 fields, according to MEDITS protocol, where not allowed: WING_OPENING, WARP_DIAMETER and VERTICAL_OPENING;
2. Check the class of the fields included in TA (e.g. numeric);
3. Check consistency between distance and duration of the haul ;
4. Check consistency between duration and time by haul;
5. Check consistency of bridles length according to MEDITS protocol;
6. Check consistency of stratum code according to the mean depth of the haul;
7. Check consistency of the hauls coordinates with the distance (difference not greater than 30%);
8. Check the consistence of the contained in specific fields with the relative allowed values;
9. Check difference between start depth and end depth (not greater than 20%) by haul;
10. Check if the values in "WING_OPENING" and "VERTICAL_OPENING" fields are in the allowed ranges according to MEDITS protocol;
11. Check the correctness of the headers ;
12. Check the presence of duplicated rows;
13. Check if the coordinates are in the GSA ;
14. Check the presence of empty fields ;
15. Check numeric range according to MEDITS protocol;
16. Check consistency of temperature by haul;
17. Check quasi-identical records;
18. Check if start depth and end depth are in the same stratum by haul;
19. Check start quadrant and end quadrant by haul;
20. Check uniqueness of valid hauls;
21. Check on the relation between shooting depth and warp length, and between warp length and wing opening;
22. Visual check of the haul positions, of the number of hauls time series and survey period by year.

2.3.2 TB table (catch data)

1. Check the class of the fields included in TB (e.g. numeric);
2. Check consistency between not null weight and not null total number;
3. Check if the total number of individuals is consistent with the sum of the individuals per sex;
4. Check if number of individuals and total weight are consistency;
5. Check correctness of species codes according to MEDITS protocol;
6. Check dictionary according to MEDITS protocol;
7. Check the correctness of the headers;
8. Check the presence of duplicated rows;
9. Check the presence of empty fields;
10. Check numeric range according to MEDITS protocol;
11. Check presence of total number and number per sex TB for species G1;
12. Check quasi-identical record in TB.

2.3.3 TC table (length and aggregated biological data)

1. Check the class of the fields included in TC (e.g. numeric);
2. Check if the length measures are reported with the correct precision;
3. Check consistency of length distribution, whether the length classes by species are included in the range reported in the DataSpecies dataset;
4. Check consistency of maturity stages according to MEDITS protocol;

5. Check consistency of maturity stages TC by the comparison with the length of smallest mature individuals reported in bibliography;
6. Check correctness of LENGTH_CLASSES_CODE according to MEDITS protocol;
7. Check correctness of number per sex in TC;
8. Check correctness of species codes according to MEDITS protocol;
9. Check dictionary according to MEDITS protocol;
10. Check the correctness of the headers;
11. Check the presence of duplicated rows;
12. Check the presence of empty fields;
13. Check numeric range according to MEDITS protocol;
14. Check presence of lengths for G1 and G2 Medits species;
15. Check quasi-identical record in TC;
16. Check about the presence of subsamples < 0.1 of the total catch;
17. Check total weight in the haul in TC.

2.3.4 Cross checks

1. Check the correctness of the number per sex in TB in case of sub-sampling in TC;
2. Check on date by haul;
3. Check presence in TA of TB hauls;
4. Check presence in TB of TA hauls;
5. Check presence in TB of TC species;
6. Check presence in TC of TB target species.

3. Technical features

The RDBqc and RoME packages have been created with the R version 4.1.2, within R studio environment (version 2022.12.0 Build 353). The roxygen2 library was used to automatically generate the function documentation in-line with code.

Devtool (R CMD check) has been used to build the package in Rstudio running all sorts of checks on the contents of an R package; this tool gives warning and error messages when it finds things that are not properly specified within the package. It also runs the examples in the .Rd files for each of the functions, as well as other automated included tests.

Before submitting a package to the R CRAN, R CMD check (including the option --as-cran) needs to be carried out to verify that in the package there are no warnings or errors and other incompatibilities with the CRAN policies.

The *dependencies* of the RDBqcpackage are: dplyr, ggplot2, rworldmap, sp, rworldxtra, pander, data.table, grDevices, magrittr, tictoc,tidyverse, fishmethods,tidyr,gridExtra, scales, outliers, sf, methods.

The *dependencies* of the RoME package are: timeDate, Stringr, ggplot2, rnaturalearth, rnaturalearthdata.

The structure followed in the two packages is showed in Figure 3.1 and is in line with the structure established by R CRAN. In R folder all the R scripts corresponding to quality check functions are contained; in “man” folder the code related to the documentation of the functions and of data examples are stored. In data folder the example data and other useful tables are contained, while in vignette folder the tutorial

that guides the user in the application of the functions is stored. In the README_files the images and support files for the help are stored.

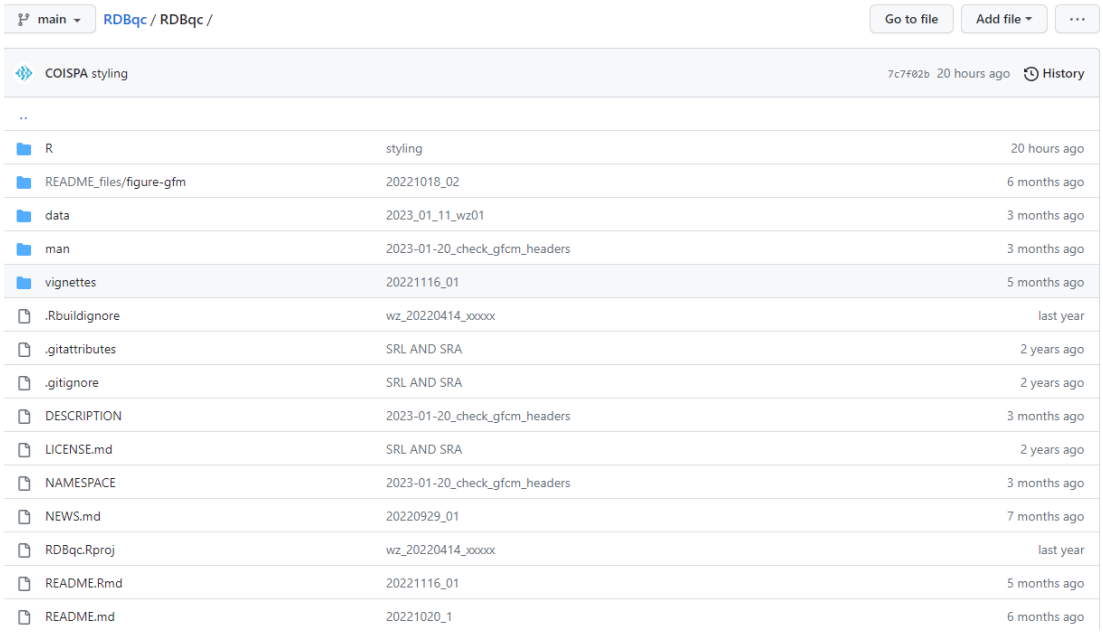

Figure 3-1 Example of structure (RDBqc) followed in the creation of RDBqc and RoME packages.

4. Testing

Several experts were involved in the testing phase of RDBqc and RoME packages. In particular, the experts were required to test the *a priori* quality checks and the *a posteriori* quality checks into two different phases:

1. **Phase 1 (from December 2021 to January 2022)**: test of the *a priori* quality checks; the experts provided a feedback, that was considered in the improving of the next version of the RDBqc package;
2. **Phase 2 (April-May 2022)**: test of the *a posteriori* quality checks, using the latest version of the RDBqc.

The vignettes of the package were provided in each phase to guide the tester in the application of the functions.

The experts tested the package on their own data and provided a feedback in the form of a detailed report, following the vignette provided with the package.

Moreover, the RDBqc package was presented and tested during the STECF EWG 22-03 “quality checking of MED & BS data and reference points”, held from the 2nd May to the 6th May 2022.During this meeting a cross check test of the package was carried out only on the functions related to the Med & BS datacall formats. The experts involved in the testing found consistent outcomes between RDBqc and the EWG 22-03 scripts. Some suggestions to homogenize the input of the different functions have been provided by the EWG. The feedback of the STECF EWG was included in the final version of the RDBqc package.

Finally, RDBqc package was tested through RDBFIS web application on the GSA 18 data in occasion of the “Med&BS RDBFIS WP5 Meetings, testing phase - 2nd WORKSHOP” held the 23rd January 2023.

RoME package has been tested on GSAs 18 and 10 (until 2016) MEDITS data. Moreover, RoME has been tested through RDBFIS web application on the MEDITS data of GSA 20 during the “Med&BS RDBFIS WP5 Meetings, testing phase - 2nd WORKSHOP” too.

5. Conclusions

In the last years, the data quality issue gained increasing importance in the EU Data Collection Framework. This is demonstrated by the fact that several projects and initiative have been devoted to the improvement of data quality both in STECF and GFCM context.

This task provided two complete, open-access and well documented tools for fishery dependent and fishery independent data in order to dramatically reduce the risk of data failure of the MS and CPs in the data submission to the four main data calls present in Mediterranean and Black Sea: MED&BS, FDI, GFCM DCRF and RCG.

In the development of this task contacts and communications with RCG and end users (STECF and GFCM) has been implemented for identifying specific needs. This communication is expected to be maintained in the future, to continue the testing of the packages, as well as their adaptation of them to specific future data needs of end-users. In this perspective the use of open access GitHub repository for storing and the development based on R language will facilitate the code adaptation and improvement, allowing the collaboration among the experts involved in the data calls.

6. References

MEDITS-Handbook. Version n. 9, 2017, MEDITS Working Group : 106 pp.

Scientific, Technical and Economic Committee for Fisheries (STECF) – Methods for supporting stock assessment in the Mediterranean (STECF-21-02). Publications Office of the European Union, Luxembourg, 2021, EUR 28359 EN, ISBN 978-92-76-40594-8, doi:10.2760/457201, JRC126125.

Scientific, Technical and Economic Committee for Fisheries (STECF) Quality checking of MED & BS data and reference points (STECF-22-03). Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/465703, JRC130288.

Spedicato M.T., Massutí E., Mérigot B., Tserpes G., Jadaud A., Relini G. 2019. The MEDITS trawl survey specifications in an ecosystem approach to fishery management. Sci. Mar. 83S1: 9-20. https://doi.org/10.3989/scimar.04915.11X.

Wilkinson, M.D., et al. 2016. The FAIR Guiding Principles for scientific data manamagement and stewardship. Scientific Data 3:160018. 9 pp. *https://doi.org/10.1038/sdata.2016.18*