Call MARE/2020/08

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

Development of the Regional Database for the Mediterranean & Black Seas

Tutorial: R markdown for automatic reporting on RCG, MED & BS, FDI and GFCM datacalls

Walter Zupa, Isabella Bitetto, IoannisChamodrakas, StefanosKavadas

Partners involved:

COISPA, HCMR

Table of contents

Introduction

Four specific R markdown automatic report files have been produced to perform both *a priori* and *a posteriori* checks to carry out data validation and quality checks respectively on detailed data and output data in different data calls required by end-users.

All the scripts useful to produce automatic reports on the above mentioned datacalls' tables are freely available on GitHub at the following repository, in the sub-folder "RMD reports": https://github.com/COISPA/RDBqc

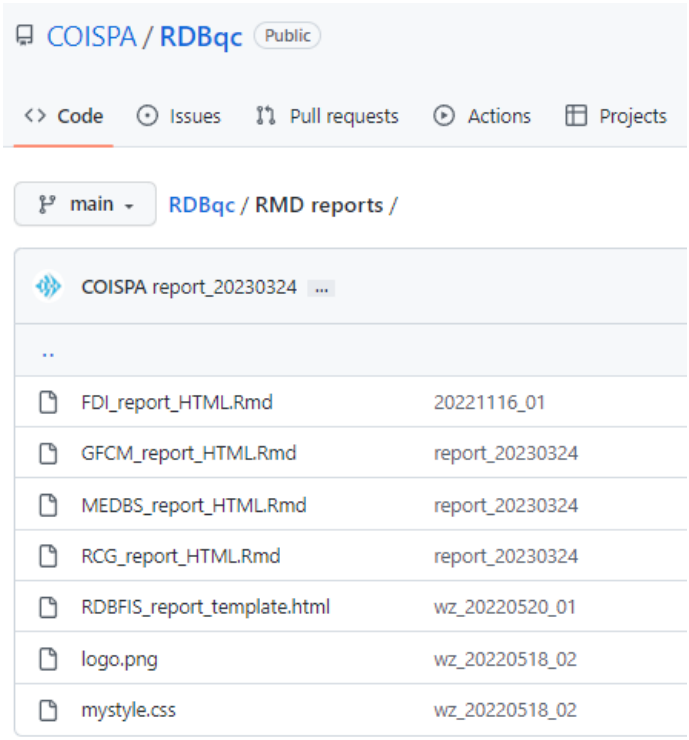

Figure 1. screenshot of the content of the GitHub repository

The folder is composed by the 4 Rmd files (Figure 1), one for each datacall:

- RCG_report_HTML.Rmd
- MEDBS_report_HTML.Rmd
- FDI_report_HTML.Rmd
- GFCM_report_HTML.Rmd

Furthermore, 3 more files (layout files) used to set the layout of the report are included in the main folder. In particular, there is a template file (RDBFIS_report_template.html), a .css style file (mystyle.css) and the logo of project (logo.png).

The Rmd file need to be used in Rstudio environment. The output of each automatic reporting files is a detailed report in HTML format of the checks carried out on the selected tables.

Environment

The Rmd file have been developed with R software (R core Team, 2022) version 4.1.2 (64-bit) and tested in Rstudio software (Rstudio Team, 2020; version 2022.12.0 Build 353) that is an integrated development environment (IDE) for R.

All the Rmd files were built to work both as standalone tools to check local data in csv format and as tools embedded in the RDBFIS database's web application, working on data imported and fed by the database. Thanks to the R multi-platform user interface, all the Rmds have been successfully tested in both Windows and Linux (Debian) platforms.

The following R packages are required to be installed in the environment for the use of the automatic reporting scripts: *RDBqc, knitr, markdown, kableExtra, dplyr, ggplot2, rworldmap, sp, rworldxtra, pander, data.table, grDevices, magrittr, tictoc, tidyverse, fishmethods, tidyr, gridExtra, outliersRDBqc, knitr, markdown, kableExtra, dplyr, ggplot2, rworldmap, sp, rworldxtra, pander, data.table, grDevices, magrittr, tictoc, tidyverse, fishmethods, tidyr, gridExtra, outliers* (Allaire et al., 2022; Auguie, 2017; Bitetto and Zupa, 2022; Daróczi and Tsegelskyi, 2021; Dowle and Srinivasan, 2021; Izrailev,2021; Komsta, 2022; Milton and Wickham, 2020; Nelson, 2021; Pebesma and Bivand, 2005; R core Team, 2022; South, 2011; South, 2012; Wickham, 2016; Wickham et al., 2019; Wickham, 2021; Wickham et al., 2021; Xie, 2022; Zhu, 2021).

Run the scripts

R (> 4.1) and Rstudio software should be already installed on the computer to run the Rmd scripts for the automatic reporting. R software can be downloaded from the CRAN website (https://cran.r-project.org/), while the Rstudio software can be downloaded from https://posit.co/products/open-source/rstudio/.

When the software is completely installed on the system, launch the Rstudio application and check the presence of the needed packages in R. Rstudio user interface (GUI) is divided in frames, and in some cases the frames are composed by tabs. Selecting the tab called "Packages" it is possible to visualise the complete list of libraries already installed in the software (Figure 2). If one or more libraries are not included in this list, press the "Install" button, in the top-left part of the frame to install the missing packages.

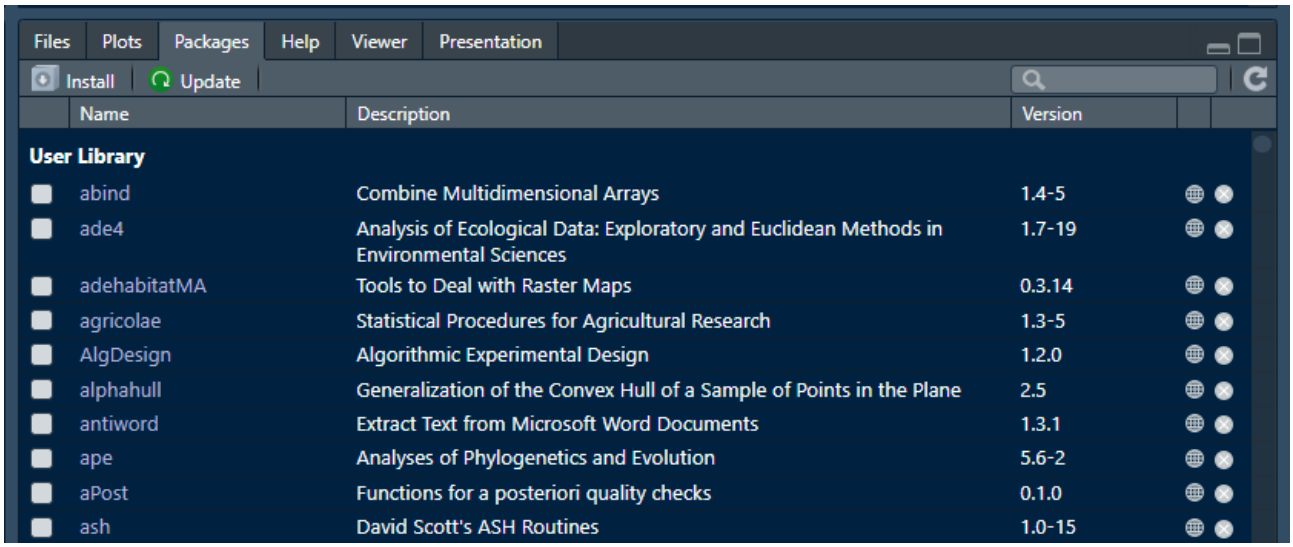

Figure 2. Screenshot of the Pachage tab in Rstudio GUI.

Once all the package are available too, the system is ready to perform the data analysis for the automatic reporting. Download the Rmd files from the GitHub repository (https://github.com/COISPA/RDBqc) and save the content of the "Rmd reports" directory in a folder in which you want to perform the analysis. R markdown will use the location of the Rmd file as the working directory.

Warning: Layout files must be located in the same folder of the Rmd files to work correctly (Figure 1).

To run the analysis and compile the report in the HTML format, press the "knitr" button in Rstudio GUI, as shown in Figure 3.

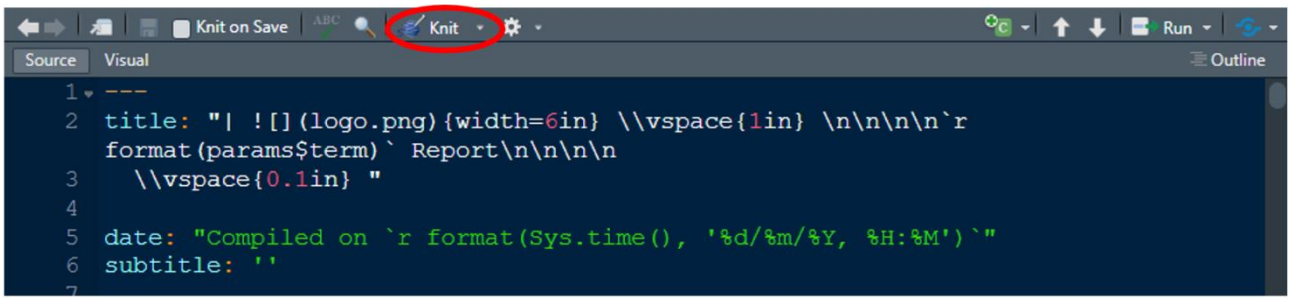

Figure 3. The position of the knitr button is indicated by the red circle

1. RCG datacall reports

Load the RCGRmd file ("RCG_report_HTML.Rmd") in Rstudio environment. Afterward, set the working directory (the one containing the Rmd files). To set the working directory, use the following line of code:

```
setwd("C:\\selected_directory_path")
```

As an alternative, select "Session" from the menu bar on the top-left part of the Rstudio window and then select "Set Working Directory" and "Choose Directory…" to set the working directory (Figure 4).

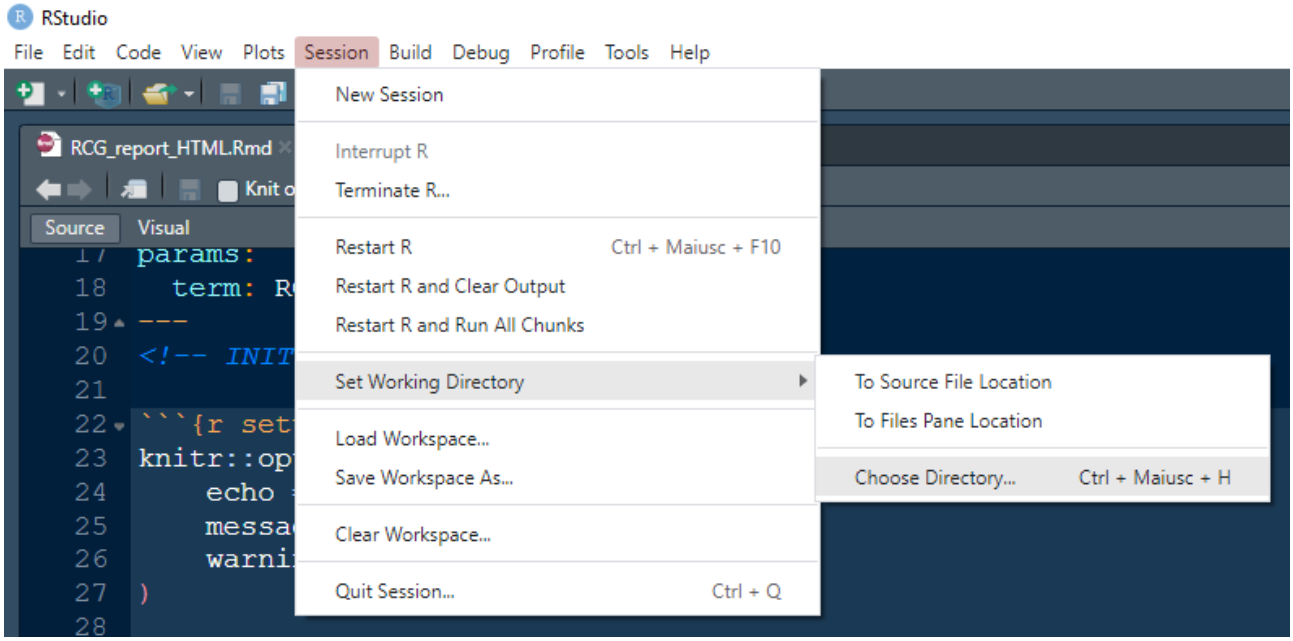

Figure 4. Screenshot of Rstudio GUI illustrating the way to set the working directory

To launch the analysis and produce the automatic reports of quality checks on the RCG datacall format the user have to modify the content of the chunk named “setup” (Figure 5), that is the only part of the Rmd code that can be modified by the user. In the case of the RCG datacall, there is the possibility to set the path for both sampling and landing data, pointing to the relative comma separated values files (csv), or set the path for at least one of the two data tables. The Rmd script will check the number of paths set by the user and will perform the analysis accordingly. In this way the user can choose to carry out the analysis either on only one table or on both of the tables. To exclude one of the two tables from the analysis the relative code line should be commented, putting the “#” symbol at the beginning of the line.

```
<!-- USER DEFINED SOURCE FILES -->

```{r setup, include=FALSE}
# reading files
# UNCOMMENT  THE FOLLOWING LINES FOR STAN-ALONE USE
# SELECT THE APPROPRIATE FILE PATHS ON THE LOCAL FOLDERS
CS <- read.csv("./sampling.csv")
CL <- read.csv("./landings.csv")


SPs <- c("Merluccius merluccius")
MS <- "ITA"
GSAs <- c("GSA99")
stages_for_immatures <- c("0", "1", "2a")
```
```

Figure 5. Screenshot of the "setup" chunk which can be modified by the final user of the RCG Rmd file to set the paths for the landing and sampling data and filters for data selection.

Furthermore, the user can set in the “setup” chunk the filters to perform the analysis only on a data subset. The user can define filters on geographical subarea (GSA) and species and set the opportune value for the

country variable (MS). To exclude the use of a given filter from data subsetting, the relative code lines should be commented with the "#" symbol.

Warning: the script runs with data from one country (MS) per time. The country code should be selected according to the GSA filter.

Some checks on sampling (CS) table related to the consistency of maturity stages require the definition of the stages to be considered as immature. The user can modify the default set of stages ("1" and "2a") modifying the content of the "stages_for_immature" variable in the chunk.

2. MED&BS datacall reports

To perform the analysis on MED & BS datacall tables, the "MEDBS_report_HTML.Rmd" file should be loaded in the Rstudio environment and the working directory (the one containing the Rmd files) set as described above in section "RCG datacall reports".

To launch the analysis and produce the automatic reports of quality checks on the MED & BSdatacall format the user have to modify the content of the chunk named "setup" (Figure 6), that is the only part of the Rmd code that can be modified by the user. In the case of the MED & BSdatacall, users have the possibility to set the path for 9 different types of tables, pointing to the relative comma separated values files (csv), or set the path for at least one of the 9 tables (catch, landing, discards, maturity at age, maturity at length, sex ratio at length, sex ratio at age, growth parameters, age-length keys). The Rmd script will check the number of paths set by the user and will perform the analysis accordingly. In this way the user can choose to carry out the analysis either on only one, some or all the tables. To exclude one of the datacall tables from the analysis the relative code line should be commented, putting the "#" symbol at the beginning.

```
<!-- USER DEFINED SOURCE FILES -->

```{r setup, include=FALSE}

# reading files
# UNCOMMENT  THE FOLLOWING LINES FOR STAND-ALONE USE
# SELECT THE APPROPRIATE FILE PATHS ON THE LOCAL FOLDERS


Catch = read.table("./catch.csv",sep=";",header=TRUE)

Land = read.table("./landings_gsa18.csv",sep=";",header=TRUE)

Disc =  read.table("./discards_gsa18.csv",sep=";",header=TRUE)

ML = read.table("./ml.csv",sep=";",header=TRUE)

MA = read.table("./ma.csv",sep=";",header=TRUE)

SL = read.table("./srl.csv",sep=";",header=TRUE)

SA = read.table("./sra.csv",sep=";",header=TRUE)

GP = read.table("./gp.csv",sep=";",header=TRUE)

ALK = read.table("./alk.csv",sep=";",header=TRUE)

SPs <- c("HKE","MUT","DPS")
MS <- "ITA"
GSAs <- "GSA 18"

```
```

Figure 6. Screenshot of the "setup" chunk which can be modified by the final user of the MED & BS Rmd file to set the paths to load the relevant tables and filters for data subsetting.

Furthermore, the user can set in the “setup” chunk the filters to perform the analysis only on a data subset. The user can define filter on geographical subarea (GSA) and species and set the opportune value for the country variable (MS). To exclude the use of a given filter from data sub setting, the relative code lines should be commented with the “#” symbol.

Warning: the script runs with data from one country (MS) per time. The country code should be selected according to the GSA filter.

3. FDI datacall reports

To perform the analysis on FDI datacall tables, the “FDI_report_HTML.Rmd” file should be loaded in the Rstudio environment and the working directory (the one containing the Rmd files) set as described above in section “RCG datacall reports”.

To launch the analysis and produce the automatic reports of quality checks on the FDI datacall format the user have to modify the content of the chunk named “setup” (Figure 7), that is the only part of the Rmd code that can be modified by the user. In the case of the FDI datacall, users have the possibility to set the path for 5 different types of tables, pointing to the relative comma separated values files (csv), or set the path for at least one of the 5 tables relevant for the Mediterranean Sea (catch, effort, landings by rectangle, effort by rectangle, capacity and fleet segment effort). The Rmd script will check the number of paths set by

the user and will perform the analysis accordingly. In this way the user can choose to carry out the analysis either on only one, some or all the tables. To exclude one of the tables from the analysis the relative code line should be commented, putting the "#" symbol at the beginning.

```
<!-- USER DEFINED SOURCE FILES -->

```{r user_setup, include=FALSE}

## reading files
## UNCOMMENT  THE FOLLOWING LINES FOR STAN-ALONE USE
## SELECT THE APPROPRIATE FILE PATHS ON THE LOCAL FOLDERS

tableA <- read.table("./fdi_a_catch.csv",sep=";",header=TRUE)
tableG <- read.table("./fdi_g_effort.csv",sep=";",header=TRUE)
tableH <- read.table("./fdi_h_spatial_landings.csv",sep=";",header=TRUE)
tableI <- read.table("./fdi_i_spatial_effort.csv",sep=";",header=TRUE)
tableJ <- read.table("./fdi_j_capacity.csv",sep=";",header=TRUE)

MS <- "ITA"
SPs = c("DPS")
GSAs <- c("GSA99")
vessel_len = "COMBINED"
fishtech = "COMBINED"
metier <- "COMBINED"



```
```

Figure 7. Screenshot of the "setup" chunk which can be modified by the final user of the FDI Rmd file to set the paths to load the relevant tables and filters for data subsetting.

Furthermore, the user can set in the "setup" chunk the filters to perform the analysis only on a data subset. The user can define filter on geographical subarea (GSA) and species and set the opportune value for the country variable (MS). Furthermore, filter on vessel length, fishing technique and metiers can be applied. To exclude the use of a given filter from data subsetting, the relative code lines should be commented with the "#" symbol.

Warning: the script runs with data from one country (MS) per time. The country code should be selected according to the GSA filter.

4. GFCM datacall reports

To perform the analysis on GFCM datacall tables, the "GFCM_report_HTML.Rmd" file should be loaded in the Rstudio environment and the working directory (the one containing the Rmd files) set as described above in section "RCG datacall reports".

To launch the analysis and produce the automatic reports of quality checks on the GFCM datacall format the user have to modify the content of the chunk named "setup" (Figure 8), that is the only part of the Rmd code that can be modified by the user. In the case of the GFCM datacall, users have the possibility to set the path for 5 different types of tables, pointing to the relative comma separated values files (csv), or set the path for at least one of the 5 tables (task II.2, task III, task VII.2, task VII.3.1, task VII.3.2). The Rmd script

will check the number of paths set by the user and will perform the analysis accordingly. In this way the user can choose to carry out the analysis either on only one, some or all the tables. To exclude one of the tables from the analysis the relative code line should be commented, putting the “#” symbol at the beginning.

```
<!-- USER DEFINED SOURCE FILES -->

```{r user_setup, include=FALSE}

# reading files
# UNCOMMENT  THE FOLLOWING LINES FOR STAND-ALONE USE
# SELECT THE APPROPRIATE FILE PATHS ON THE LOCAL FOLDERS

T_ii2 <- read.csv("./dc_dcrf_task_ii2_catch.csv", sep=";")

T_iii <- read.csv("./dc_dcrf_task_iii_incidental_catch.csv", sep=";")

T_vii2 <- read.csv("./dc_dcrf_task_vii2_length_data.csv", sep=";")

T_vii31 <- read.csv("./dc_dcrf_task_vii31_size_1st_matur.csv", sep=";")

T_vii32 <- read.csv("./dc_dcrf_task_vii32_maturity_data.csv", sep=";")


MS <- "ITA"
SPs = c("HKE","DPS")
GSAs <- c(99)
segments = "COMBINED"
```
```

Figure 8. Screenshot of the "setup" chunk which can be modified by the final user of the GFCM Rmd file to set the paths to load the relevant tables and filters for data subsetting.

Furthermore, the user can set in the “setup” chunk the filters to perform the analysis only on a data subset. The user can define filter on geographical subarea (GSA) and species (SPs)and set the opportune value for the country variable (MS). Furthermore, a filter on fleet segment can be applied. To exclude the use of a given filter from data subsetting, the relative code lines should be commented with the “#” symbol.

Warning: the script runs with data from one country (MS) per time. The country code should be selected according to the GSA filter.

References

Allaire,J.J., Xie, Y., McPherson, J., Luraschi, J.,Ushey, K., Atkins, A., Wickham, H.,Cheng, J., Chang, W., Iannone, R. (2022). rmarkdown: Dynamic Documents for R. R package version 2.17. URL https://rmarkdown.rstudio.com

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3.

Bitetto I., Zupa W. (2022). RDBqc: Quality check functions for RDBFIS. R package version 0.0.14.

Daróczi, G., Tsegelskyi, R. (2021). pander: An R 'Pandoc' Writer. R package version 0.6.4. https://CRAN.R-project.org/package=pander

Dowle, M, Srinivasan, A. (2021). data.table: Extension of `data.frame`. R package version 1.14.2. https://CRAN.R-project.org/package=data.table

Gruber, J. (2004). https://daringfireball.net/projects/markdown/

Izrailev, S. (2021). tictoc: Functions for Timing R Scripts, as Well as Implementations of Stack and List Structures. R package version 1.0.1. https://CRAN.R-project.org/package=tictoc

Komsta, L. (2022). outliers: Tests for Outliers. R package version 0.15. https://CRAN.R-project.org/package=outliers

Milton, S. B., Wickham, H. (2020). magrittr: A Forward-Pipe Operator for R. R package version 2.0.1. https://CRAN.R-project.org/package=magrittr

Nelson, G. A. (2021). fishmethods: Fishery Science Methods and Models. R package version 1.11-2. https://CRAN.R-project.org/package=fishmethods

Pebesma, E. J., Bivand, R.S. (2005). Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

South, A. (2011) rworldmap: A New R package for Mapping Global Data. The R Journal Vol. 3/1 : 35-43.

South, A. (2012). rworldxtra: Country boundaries at high resolution. R package version 1.01. https://CRAN.R-project.org/package=rworldxtra

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, H. (2021). tidyr: Tidy Messy Data. R package version 1.1.4. https://CRAN.R-project.org/package=tidyr

Wickham, H.(2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wickham, H., Romain, F., Lionel, H., Kirill, M. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Xie, Y. (2022). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.40.

Zhu, H (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4.