Open and Reproducible Fisheries Science

Standardized workflows at ICES and FAO

Arni Magnusson[1], **Colin Millar**[2], **Rishi Sharma**[3]

[1] Pacific Community (SPC), Nouméa, New Caledonia
[2] International Council for the Exploration of the Sea (ICES), Copenhagen, Denmark
[3] UN Food and Agriculture Organization (FAO), Rome, Italy

CAPAM Good Practices Workshop
Rome, 24 Oct 2022

Overview

Why *repeatability, institutional memory, reviewability, scientific method, interregional research, dissemination, collaboration, traceability, credibility*

Open *scripts, data, software*

Reproducible *standardized sequential R scripts, version control*

Infrastructure *2021 UN Recommendation on Open Science, working group, GitHub, TAF, data management, ICES, FAO, GFCM, SPC*

Recommendations *relative paths, dependencies, 1st and 2nd class scripts, complete workflow, data preparation, partially open*

Why open and reproducible

Peer reviewability and reproducibility are **core principles** underpinning the **scientific method**

→ Reproducibility distinguishes between arbitrary analyses and science

→ Some assessments are more reviewable than others

Easy to pick up a stock assessment from a previous year and run an **update assessment**

→ Especially important when a new scientist is doing the assessment

Why open and reproducible (cont)

Easy to modify model settings and rerun the **entire workflow**

→ Allows more thorough analyses, exploration, improvements

Share data that others can use

→ Dissemination, machine and human readable

→ Interregional research, collaboration

Traceability

→ Credibility, buy-in

Compare to 1990s:

litdb, photocopy

buying software

typing numbers

Open

Scripts	GitHub
Data	Static HTML
	GitHub
	Data warehouse
	Web services
Software	GitHub
	Releases

How Open?

Not provided online

Sensitive data

Requires login

Available by request

Part private, part open

Hard to find

Fully open

Easy to find and browse

Reproducible Analysis

Can be run on any computer

- By different people
- On different operating systems
- In different software environments

Can be run later

- Next week
- Next year
- 5–10 years from now

Can be modified and rerun

- With different data
- With different model settings

How Reproducible?

A gradient from low $\rightarrow$ high **quality of science**, in terms of reproducibility:

1. Here's the management advice – trust me, I did the math
2. I used the model published in this paper and here are the data tables and results
3. I used these exact equations and preprocessed the data in this manner
4. Here are some scripts that give the general idea
5. Here are scripts that run on my computer, as a complete workflow without errors
6. Here are scripts that should run on your computer, along with all input files and software dependencies
7. I've cleaned up the directory to include only files required to run the core analysis, tested on another computer, with exact instructions on how to run
8. Adopted a standard reproducible format for the analysis

How to Make an Analysis Reproducible

R scripts Relative paths (data/catch.dat)
Can be run from command line: `Rscript myscript.R`
Manageable size

General structure
1. Load packages
2. Read files
3. Do the work
4. Write files

Standardize further One script prepares data
Another script runs the core analysis
Third script gathers the results
Fourth script produces plots and formatted tables for report
$\Rightarrow$ Every script is self-contained, reading files from previous steps
$\Rightarrow$ Every analysis is structured the same, anyone can pick up and run

Transparent Assessment Framework (TAF)

Supports open and reproducible science

More than anything, an **agreed way** to work with R scripts

Launched in 2018

Developed by ICES, used around the world

taf.ices.dk


IRAQ ROOM
قاعة العراق
FISHFORUM
Bridging science, seas, people

Transparent Assessment Framework (TAF)

TAF applications

- Around 30 ICES stock assessments each year
- ICES survey indices
- ICES catch at age
- ICES fisheries overviews
- FAO SOFIA – under development

R packages TAF, icesTAF, SOFIA

Version control – software and data

Data provenance – who, what, where

SOFIA-TAF

Standardized structure to organize the SOFIA analyses

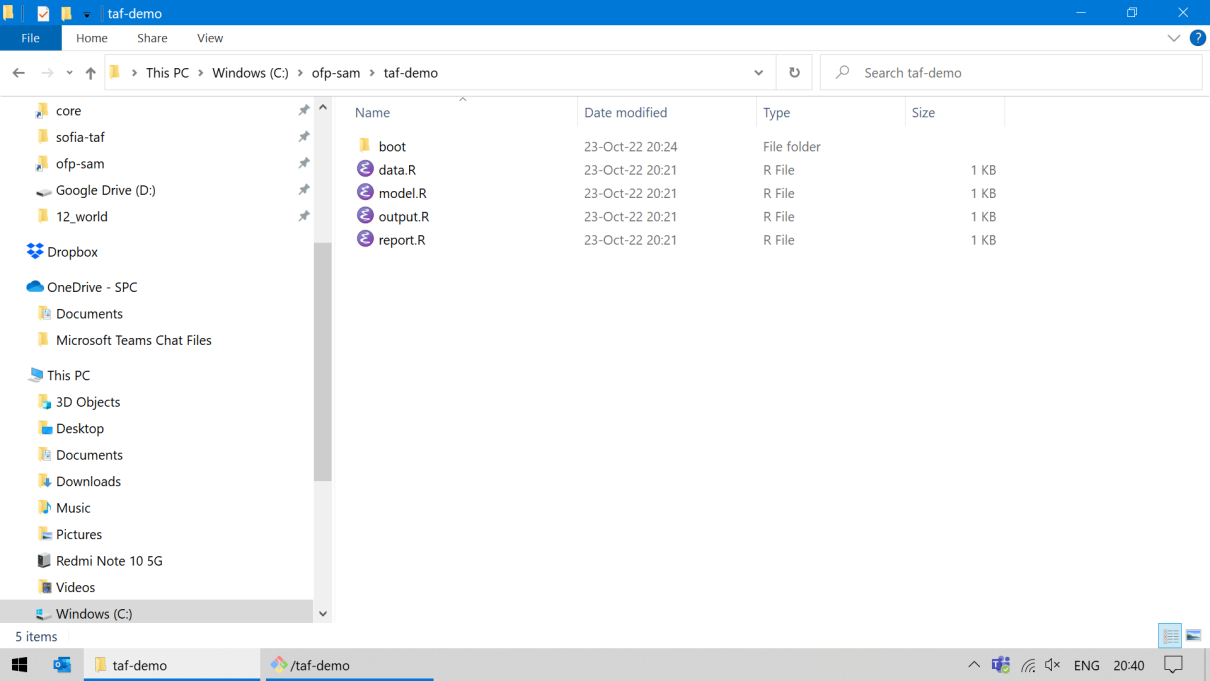
All the fisheries in the world

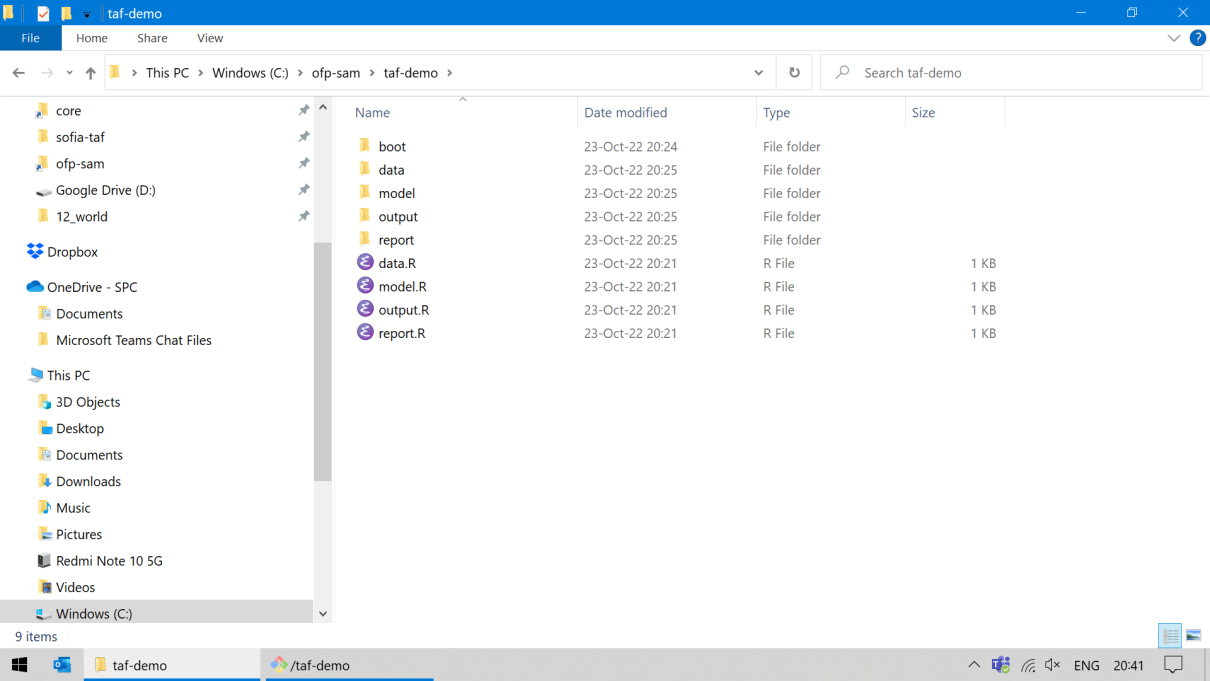
Converted from monolithic R Markdown to modular scripts

Tiers 1, 2, and 3

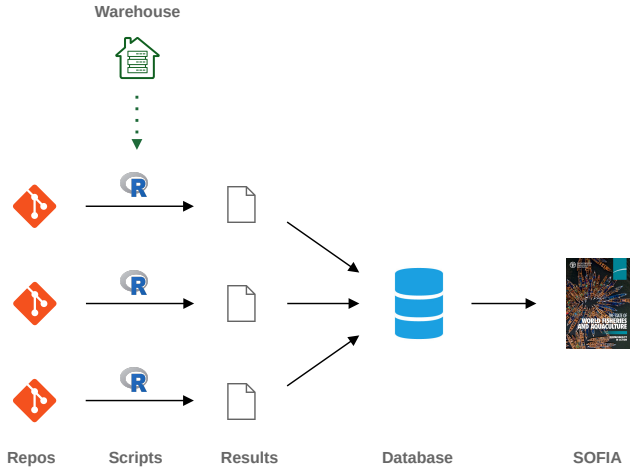
Ongoing development at github.com/sofia-taf

R package SOFIA, one place to make changes, affecting all the analyses


taf-demo
File
Home
Share
View
This PC
Windows (C:)
ofp-sam
taf-demo
Search taf-demo
core
sofia-taf
ofp-sam
Google Drive (D:)
12_world
Dropbox
OneDrive - SPC
Documents
Microsoft Teams Chat Files
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Redmi Note 10 5G
Videos
Windows (C:)
Name
Date modified
Type
Size
boot
23-Oct-22 20:24
File folder
data.R
23-Oct-22 20:21
R File
1 KB
model.R
23-Oct-22 20:21
R File
1 KB
output.R
23-Oct-22 20:21
R File
1 KB
report.R
23-Oct-22 20:21
R File
1 KB
5 items
taf-demo
/taf-demo
ENG
20:40


taf-demo
File
Home
Share
View
This PC
Windows (C:)
ofp-sam
taf-demo
Search taf-demo
core
sofia-taf
ofp-sam
Google Drive (D:)
12_world
Dropbox
OneDrive - SPC
Documents
Microsoft Teams Chat Files
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Redmi Note 10 5G
Videos
Windows (C:)
Name
Date modified
Type
Size
boot 23-Oct-22 20:24 File folder
data 23-Oct-22 20:25 File folder
model 23-Oct-22 20:25 File folder
output 23-Oct-22 20:25 File folder
report 23-Oct-22 20:25 File folder
data.R 23-Oct-22 20:21 R File 1 KB
model.R 23-Oct-22 20:21 R File 1 KB
output.R 23-Oct-22 20:21 R File 1 KB
report.R 23-Oct-22 20:21 R File 1 KB
9 items
taf-demo
/taf-demo
ENG
20:41

SOFIA-TAF



Open Science Infrastructures

Some examples from fisheries

2018	ICES	ICES-TAF
2021	FAO	GFCM STAR
2022	FAO	SOFIA-TAF
2023	SPC	Std Repos

General

2021 UN Recommendation on Open Science

$\rightarrow$ adopted by 193 member states, provides principles and norms

2022 UN Working Group on Open Science Infrastructures

Discussion

There are two types of scripts

1. Script that **runs**
   relatively short, does one part of the workflow, as reflected by its filename,
   $\Rightarrow$ *to use this script: run it*

2. Script that is **essentially a notebook**
   longer, does many parts of the workflow,
   not really intended to run completely from start to finish,
   $\Rightarrow$ *to use this script: open and run selected blocks of code*

It's useful to clearly distinguish between these types of scripts (1st and 2nd class) and consider what is most practical for a given project

Discussion

Data preparation is a large part of the stock assessment work

- Indices *survey, cpue*
- Catch *in tonnes, age composition, size composition*
- Life history *maturity, weightj*
- Tags *releases, recaptures*
- etc.

The benefits of **scripting the data preparation** as reproducible workflows can be at least as important and benefitial as scripting the model run

Recommendations

Project organization

1. Organize analyses in GitHub repos, preferably open (at least scripts)
2. Divide the work into separate scripts of manageable size, with descriptive names
3. Make each script run by itself: read files, do stuff, write files
4. Consider using a common structure to organize projects

Within each script

1. Write the script so it will run on any computer
2. Avoid using setwd() use alt-sws in RStudio
3. Use relative paths
4. Use few dependencies Chuck Norris style

Transparency in Fisheries Management

Transparent = open and reproducible
as a result, reviewable and traceable

A growing question in all fisheries around the world:
⇒ **Is the management of this stock based on open and reproducible science?**

If not, which criteria are still missing?

Overview

Why *repeatability, institutional memory, reviewability, scientific method, interregional research, dissemination, collaboration, traceability, credibility*

Open *scripts, data, software*

Reproducible *standardized sequential R scripts, version control*

Infrastructure *2021 UN Recommendation on Open Science, working group, GitHub, TAF, data management, ICES, FAO, GFCM, SPC*

Recommendations *relative paths, dependencies, 1st and 2nd class scripts, complete workflow, data preparation, partially open*