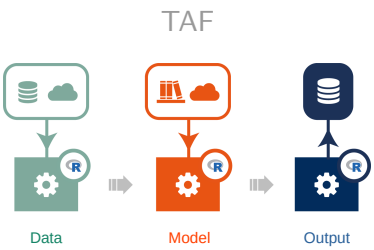
Open and Reproducible Fisheries Science

Arni Magnusson

GitHub



SPC Pre-Assessment Workshop (PAW)
Noumea, 31 March 2022

Overview

Open *scripts, data, software*

Reproducible *standardized sequential R scripts, version control*

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

Tools *GitHub, TAF*

SPC *tools, internal workflows, stock assessments, reviews, MSE*

Open

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

Reproducible Analysis

Pacific Community
Communauté du Pacifique

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 data preparation
- With different model options

Why

1. Repeatability
2. Institutional memory
3. Reviewability
4. Scientific method
5. Interregional research
6. Dissemination
7. Collaboration
8. Traceability
9. Credibility

How to Make an Analysis Reproducible

Pacific Community
Communauté du Pacifique

R scripts
Relative paths (data/catches.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

Pacific Community
Communauté du Pacifique

GitHub and TAF

What is GitHub

Free website where people can make things **available for download**:

Software	*stock assessment models, R packages*
Analyses	*R scripts, aggregated data tables*

Also a **collaborative work** environment:

Software development	*produce software, distribute*
Expert groups	*do science, share tools*

Examples `https://github.com/PacificCommunity`
`https://github.com/arni-magnusson`

Pacific Community (SPC)

Pacific Community (SPC) Official GitHub Organization

New Caledonia http://www.spc.int/ hostmaster@spc.int

Follow

Repositories

Find a repository... Type Language Sort New

ofp-sam-yft-review Public
Yellowfin tuna assessment review
0 0 0 0 Updated 3 hours ago

OFP-Ikamoana Private
Jupyter Notebook 1 1 0 0 Updated 8 hours ago

OFP-FEMA-BioDaSys Private
Local front-end for the Biological Database System (Tissue Bank)
C# 0 0 6 0 Updated 11 hours ago

tuf_common Private
Common Tuf code shared between Tuf instances like Tufman2
C# 0 0 0 0 Updated 15 hours ago

tufman2 Private
C# 0 1 11 (1 issue needs help) 0 Updated 16 hours ago

People

View all

Top languages

R C# JavaScript HTML TypeScript

What is GitHub

Pacific Community
Communauté du Pacifique

Version control

- Backup of previous editions
- Can always go back in history
- See who changed what, when, and why

Access control

- Public/private projects
- Read/write access for each user

Social network

- Follow projects and colleagues
- Send comments and suggestions
- Makes work cool and fun

Browse and download files

- Easy

Upload and edit files

- Quite technical

Transparent Assessment Framework (TAF)

ICES TAF page: https://taf.ices.dk

SPC TAF demo: https://github.com/PacificCommunity/taf-demo

CRAN package: https://cran.r-project.org/package=TAF

Strictly speaking, the TAF package is not required to write your analysis in TAF format

More than anything, TAF is an

$\Rightarrow$ **Agreed way to work**

How to Make an Analysis Reproducible

Pacific Community
Communauté du Pacifique

R scripts
Relative paths (data/catches.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)

Pacific Community
Communauté du Pacifique

R scripts Relative paths (data/catches.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)

Pacific Community
Communauté du Pacifique

TAF applications

- Hundreds of ICES stock assessments
- ICES survey indices
- ICES catch at age
- ICES fisheries overviews
- FAO SOFIA (State of World Fisheries and Aquaculture) – *under development*

TAF and **icesTAF** R packages

Version control – software and data

Data provenance – who, what, where

How Reproducible?

A gradient from low → 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

Non-reproducible results are not accepted in fields like climate research and medical research

Reproducibility distinguishes between arbitrary analyses and science

Transparency in Fisheries Management

Pacific Community
Communauté du Pacifique

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?

SPC

Tools *Multifan-CL, R packages, MFCL-Viewer (Java)*

Internal workflows *data preparation, stepwise development, diagnostic model run, model grid, plots*

Stock assessments *zip file containing full model grid*

Reviews *documents, analyses*

MSE *shiny apps*

Yellowfin Tuna Review

https://github.com/PacificCommunity/ofp-sam-yft-review