

Introduction to R and computing for Quantitative Fisheries Science



Iago MOSQUEIRA
Maritime Affairs Unit - IPSC
European Commission
Joint Research Center

Why programming?

“Can one be a good data analyst without being a half-good programmer? The short answer to that is, ‘No’. The long answer to that is, ‘No!’.”

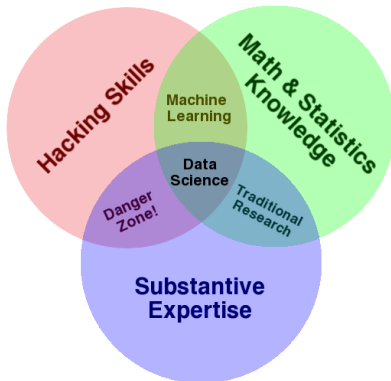
– Frank Harrell, 1999 S-PLUS User Conference, New Orleans (October 1999)

But this should be easy

“Managing fisheries is hard: it’s like managing a forest, in which the trees are invisible and keep moving around”

- Professor John Shepherd

Data analyst



What is R



- Data analysis and statistics environment
- Interpreted computer language
- Open-source software project
- Active community of developers and practitioners
- Current version: 3.1.0 (2014-04-10) – “Spring Dance”

Why R?

R is

- awesome
- free (both speech and beer)
- a community
- the *lingua franca*
- a language and an environment
- integrates with other tools
- a way to reproducible research

OSS = Peer review

Previous Next 1 (1 of 2) Fit Page Width

POLICYFORUM

COMPUTATIONAL SCIENCE

Troubling Trends in Scientific Software Use

Lucas N. Joppa,^{1*} Greg McInerney,¹² Richard Harper,¹ Lara Salido,³ Kenji Takeda,¹ Kenton O'Hara,¹ David Gavaghan,¹ Stephen Emmott¹

Software pervades every domain of science (1–3), perhaps nowhere more decisively than in modeling. In key scientific areas of great societal importance, models and the software that implement them define both how science is done and what science is done (4, 5). Across all science, this dependence has led to concerns around the need for open access to software (6, 7), centered on the reproducibility of research (1, 8–10). From fields such as high-performance computing, we learn key insights and best practices for how to develop, standardize, and implement software (11). Open and systematic approaches to the development of software are essential for all sciences. But for many scientists this is not sufficient. We

across all disciplines that are dependent upon a computational approach.

Surveying Species Distribution Modelers

We surveyed scientists across a single domain, species distribution modeling (SDM) (15) [see supplementary materials for details]. This strategic targeting separates our analysis from previous efforts in important ways, allowing an analysis spanning computational skill sets, while addressing the interplay between models and computation. Our ~400 respondents ranged from those who “find it difficult to use software” to those “very experienced and very technical.” Asking people to first identify with a scientific domain and addressing models and software through that

used “click-and-run” software with easy-to-manipulate user interfaces and dropped to 11% for those who used “syntax-driven” platforms. Further, 7, 9, and 18% of scientists cited “the developer is well-respected,” “personal recommendation,” and “recommendation from a close colleague,” respectively, as reasons for using software. Only 8% claimed they had validated software against other methods as a primary reason for choice; 79% expressed a desire to learn additional software and programming skills.

Many of these scientists rely on the fact that the software has appeared in a peer-reviewed article, recommendations, and personal opinion, as their reason for adopting software. This is scientifically misplaced, as the software code used to conduct the science

“Blind trust” is dangerous when choosing software to support research.

ag.org on May 17, 2013

GPL v3.0

GNU General Public License

From Wikipedia, the free encyclopedia

"*GPL*" *redirects here. For other uses, see [GPL \(disambiguation\)](#).*

The **GNU General Public License** (**GNU GPL** or simply **GPL**) is the most widely used^[5] [free software license](#). It was originally written by [Richard Stallman](#) for the [GNU Project](#).

The GPL is the first [copyleft](#) license for general use, which means that derived works can only be distributed under the same license terms. Under this philosophy, the GPL grants the recipients of a computer program the rights of the [free software definition](#) and uses copyleft to ensure the freedoms are preserved, even when the work is changed or added to. This is in distinction to [permissive free software licenses](#), of which the [BSD licenses](#) are the standard examples.

Contents [\[hide\]](#)

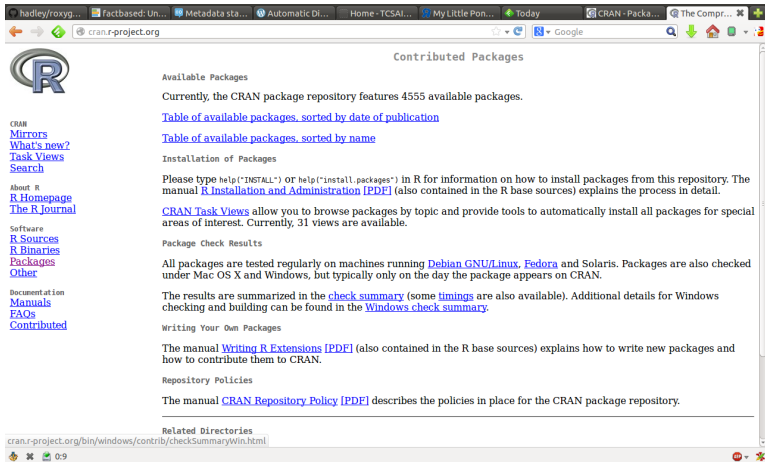
- 1 History
- 2 Versions
 - 2.1 Version 1
 - 2.2 Version 2
 - 2.3 Version 3
- 3 Terms and conditions
 - 3.1 Copyleft
- 4 Licensing and contractual issues
- 5 Copyright holders
- 6 Linking and derived works
 - 6.1 Libraries
 - 6.2 Communicating and bundling with non-GPL programs
- 7 The GPL in court
- 8 Compatibility and multi-licensing
 - 8.1 Multi-licensing
- 9 Adoption
- 10 Use for text and other media

GNU General Public License



Author	Free Software Foundation
Version	3
Publisher	Free Software Foundation, Inc.
Published	29 June 2007
DFSG compatible	Yes ^[1]
FSF approved	Yes ^[2]
OSI approved	Yes ^[3]
Copyleft	Yes ^{[2][4]}
Linking from code with a different license	No (except for linking GNU AGPLv3 with GNU GPLv3 – see section)
Website	www.gnu.org/licenses

CRAN



Contributed Packages

Available Packages

Currently, the CRAN package repository features 4555 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The manual [R Installation and Administration \[PDF\]](#) (also contained in the R base sources) explains the process in detail.

[CRAN Task Views](#) allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 31 views are available.

Package Check Results

All packages are tested regularly on machines running [Debian GNU/Linux](#), [Fedora](#) and Solaris. Packages are also checked under Mac OS X and Windows, but typically only on the day the package appears on CRAN.

The results are summarized in the [check summary](#) (some [timings](#) are also available). Additional details for Windows checking and building can be found in the [Windows check summary](#).

Writing Your Own Packages

The manual [Writing R Extensions \[PDF\]](#) (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.

Repository Policies

The manual [CRAN Repository Policy \[PDF\]](#) describes the policies in place for the CRAN package repository.

Related Directories

[cran.r-project.org/bin/windows/contrib/checkSummaryWin.html](#)

CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

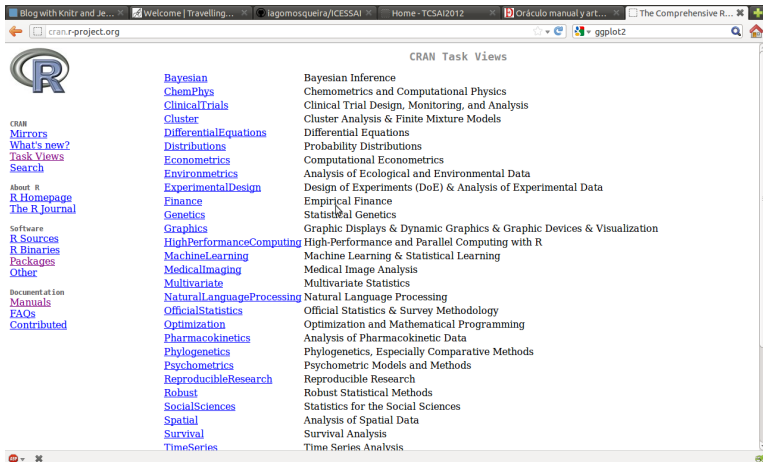
Documentation

[Manuals](#)

[FAQs](#)

[Contributed](#)

Task views



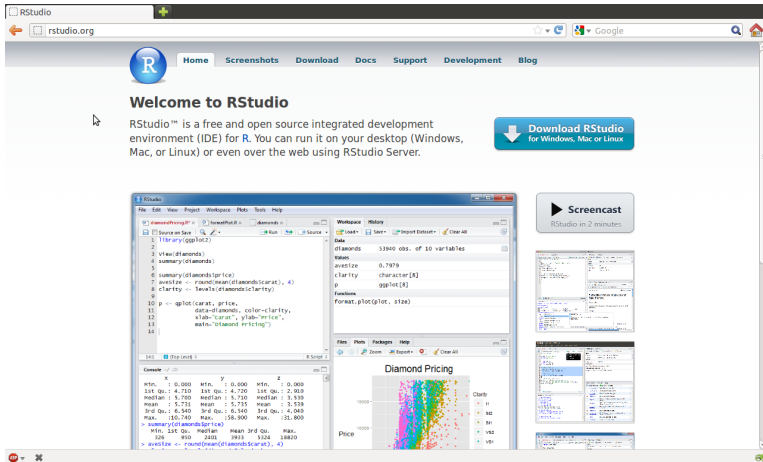
The screenshot shows a web browser window displaying the CRAN Task Views page. The browser's address bar shows 'cran.r-project.org'. The page features the R logo on the left and a list of task views in the center. The task views are organized into two columns. The left column lists various task views, and the right column lists the corresponding R packages. The task views include Bayesian Inference, Chemometrics and Computational Physics, Clinical Trial Design, Monitoring, and Analysis, Cluster Analysis & Finite Mixture Models, Differential Equations, Probability Distributions, Computational Econometrics, Analysis of Ecological and Environmental Data, Design of Experiments (DoE) & Analysis of Experimental Data, Empirical Finance, Statistical Genetics, Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization, High-Performance and Parallel Computing with R, Machine Learning & Statistical Learning, Medical Image Analysis, Multivariate Statistics, Natural Language Processing, Official Statistics & Survey Methodology, Optimization and Mathematical Programming, Analysis of Pharmacokinetic Data, Phylogenetics, Especially Comparative Methods, Psychometric Models and Methods, Reproducible Research, Robust Statistical Methods, Statistics for the Social Sciences, Analysis of Spatial Data, Survival Analysis, and Time Series Analysis.

CRAN Task Views

Bayesian
ChemPhys
ClinicalTrials
Cluster
DifferentialEquations
Distributions
Econometrics
Environmetrics
ExperimentalDesign
Finance
Genetics
Graphics
HighPerformanceComputing
MachineLearning
MedicalImaging
Multivariate
NaturalLanguageProcessing
OfficialStatistics
Optimization
Pharmacokinetics
Phylogenetics
Psychometrics
ReproducibleResearch
Robust
SocialSciences
Spatial
Survival
TimeSeries

Bayesian Inference
Chemometrics and Computational Physics
Clinical Trial Design, Monitoring, and Analysis
Cluster Analysis & Finite Mixture Models
Differential Equations
Probability Distributions
Computational Econometrics
Analysis of Ecological and Environmental Data
Design of Experiments (DoE) & Analysis of Experimental Data
Empirical Finance
Statistical Genetics
Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
High-Performance and Parallel Computing with R
Machine Learning & Statistical Learning
Medical Image Analysis
Multivariate Statistics
Natural Language Processing
Official Statistics & Survey Methodology
Optimization and Mathematical Programming
Analysis of Pharmacokinetic Data
Phylogenetics, Especially Comparative Methods
Psychometric Models and Methods
Reproducible Research
Robust Statistical Methods
Statistics for the Social Sciences
Analysis of Spatial Data
Survival Analysis
Time Series Analysis

RStudio



The image shows the RStudio website interface. At the top, there's a navigation bar with links: Home, Screenshots, Download, Docs, Support, Development, and Blog. Below this is a large "Welcome to RStudio" section. A mouse cursor points to the text: "RStudio™ is a free and open source integrated development environment (IDE) for R. You can run it on your desktop (Windows, Mac, or Linux) or even over the web using RStudio Server." To the right of this text is a blue button that says "Download RStudio for Windows, Mac or Linux".

Below the welcome section is a preview of the RStudio IDE. The IDE window is titled "RStudio" and shows a script editor with R code for loading the 'ggplot2' library, viewing the 'diamonds' dataset, and creating a scatter plot of 'price' vs 'carat' colored by 'clarity'. The console shows the output of these commands, including a summary of the 'diamonds' dataset and the execution of 'ggplot()' and 'geom_point()'. The 'Workspace' pane on the right shows the loaded objects: 'diamonds' (53940 obs. of 10 variables), 'aveprice' (0.7076), 'clarity' (character [8]), and 'p' (ggplot [8]). The 'Files' pane shows the 'Diamond Pricing' project folder. A 'Screencast' button is also visible next to the IDE preview.

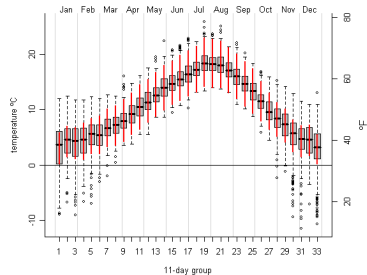
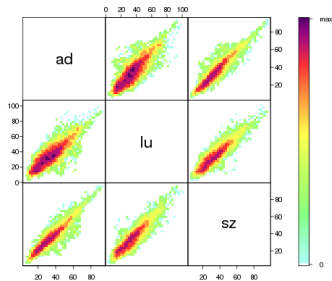
Basic features

- Numerous procedures (algebra, matrix, stats)
- Named storage (everything is an object)
- Functions
- Classes and methods (S3, S4)
- Special values (NA, NaN, Inf, NULL)
- Logical objects and boolean algebra
- `basic_features.R`

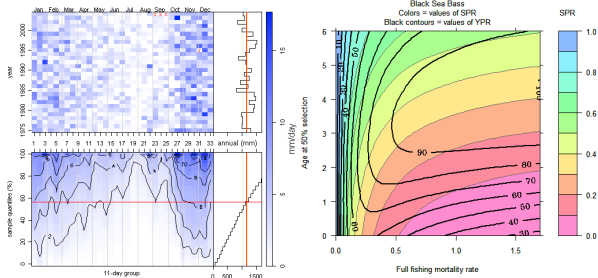
What else can it do?

- Data handling and storage
- Matrix algebra
- Regular expressions
- Statistics!
- OOP
- Programming
- Graphics

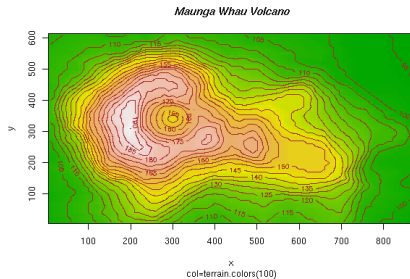
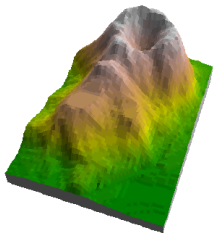
Eye candy



Eye candy



Eye candy



What doesn't it do

- No DB, but connections (SQL, NoSQL, Spreadsheets)
- No GUI, but IDE & GUI toolsets - CLI
- Slow, but C/C++, HPC
- No commercial support, but community
- Think for you

Help!

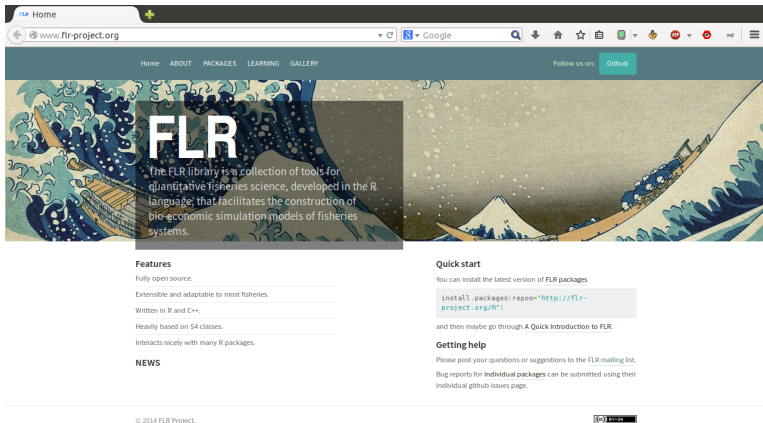


- Help for each function and data type
- ?mean
- ??mean
- ?help
- <http://rseek.org>
- stackoverflow, <http://stackoverflow.com/questions/tagged/r>
- Mailing lists

FLR

- Stock assessment and provision of management advice
 - Well tested, robust methods
 - Open to detailed inspection
- Data and model validation through simulation
- Risk analysis
- Capacity development & education
- Promote collaboration and openness in quantitative fisheries science
- Support the development of new models and methods
 - Extensible toolset
 - Links to other tools (ADMB, BUGS, ...)

flr-project.org



The screenshot shows the homepage of the flr-project.org website. The browser address bar displays 'www.flr-project.org'. The website has a dark blue header with navigation links: Home, ABOUT, PACKAGES, LEARNING, and GALLERY. A 'Follow us on: Github' button is also present. The main content area features a large illustration of a traditional Japanese boat on stylized waves. Overlaid on this is the text 'FLR' in large white letters, followed by a description: 'The FLR library is a collection of tools for quantitative fisheries science, developed in the R language; that facilitates the construction of bio-economic simulation models of fisheries systems.' Below this, there are sections for 'Features', 'Quick start', and 'Getting help'. The 'Features' section lists: 'Fully open source.', 'Extensible and adaptable to most fisheries.', 'Written in R and C++.', 'Heavily based on S4 classes.', and 'Interacts nicely with many R packages.' The 'Quick start' section provides instructions on installing the latest version of FLR packages using the command: `install.packages(repo="http://flr-project.org/R")`, and mentions a 'Quick Introduction to FLR'. The 'Getting help' section encourages users to post questions or suggestions to the 'FLR mailing list' and bug reports for individual packages to be submitted using the individual github issues page. At the bottom, there is a copyright notice '© 2014 FLR Project.' and a Creative Commons BY-NC logo.

Home

www.flr-project.org

Google

Home ABOUT PACKAGES LEARNING GALLERY

Follow us on: Github

FLR

The FLR library is a collection of tools for quantitative fisheries science, developed in the R language; that facilitates the construction of bio-economic simulation models of fisheries systems.

Features

- Fully open source.
- Extensible and adaptable to most fisheries.
- Written in R and C++.
- Heavily based on S4 classes.
- Interacts nicely with many R packages.

NEWS

Quick start

You can install the latest version of FLR packages

```
install.packages(repo="http://flr-project.org/R")
```

and then maybe go through [A Quick Introduction to FLR](#).

Getting help

Please post your questions or suggestions to the [FLR mailing list](#).

Bug reports for [individual packages](#) can be submitted using their individual github issues page.

© 2014 FLR Project.

CC BY-NC

Tools of the trade

- Version Control Systems
- Editors & IDEs
- Literate Programming
- Validation, Verification and Testing (VV&T)
- **Reproducible Research**

Sexy data analysis



Setting up R & RStudio

- <http://cran.r-project.org>
- <http://rstudio.org>