

Reproducible Research with R

D G Rossiter

Section of Soil & Crop Sciences, Cornell University

ISRIC-World Soil Information

南京师范大学地理科学学院

中国科学院南京土壤研究所

November 20, 2019

Outline

- 1 Reproducible Research
- 2 The R Project
- 3 Some simple R
- 4 Literate Data Analysis
- 5 The R ecosystem
- 6 Final words

Outline

- 1 Reproducible Research
- 2 The R Project
- 3 Some simple R
- 4 Literate Data Analysis
- 5 The R ecosystem
- 6 Final words

Reproducible research 可再现的研究

Definition: “Research papers with accompanying software tools that allow the reader to directly reproduce the results and employ the computational methods that are presented in the research paper.”¹

¹Gentleman, R., & Lang, D. T. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1), 1–23. <https://doi.org/10.1198/106186007X178663>

Why reproducible research?

- **Readers** can both **verify** and **adapt** the claims in the paper
 - no need to trust that the author has performed computations correctly
 - no hidden assumptions
- **Authors** can **reproduce** the results in the future – e.g., when a study is extended to a new area
- **Authors** can **adapt** the document to a different medium – journal paper, web page, tutorial . . .
- In-text computations, figures, tables, can be **recalculated** if the data changes
 - e.g., reading in a different climate dataset each day and generating a report

Elements of reproducible research

- 1 **Datasets** (tabular, geographic, time-series ...)
- 2 **Computer code**
- 3 **Text** explaining **processing**
- 4 **Text** showing **results** – generated by the computer code
- 5 **Graphics** showing **results** – generated by the computer code
- 6 **Text** discussing **results**

Example reproducible research document

Spatio-temporal analysis and interpolation of PM₁₀ measurements in Europe

Task 1012 - Subtask 4

ETC/ACM Technical Paper 2011/10

Benedikt Gröler, Lydia Gerharz, Edzer Pebesma

ben.graeler@uni-muenster.de

November 2011

Including Erratum

(dd. 18 March 2013)

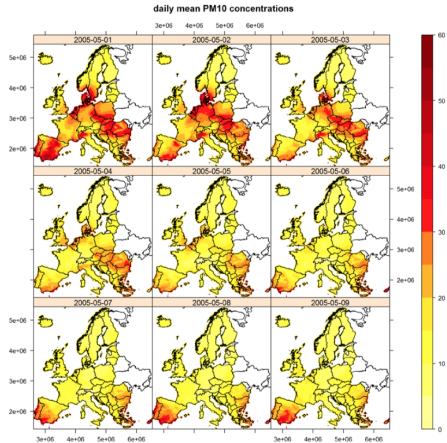


Figure 7: Interpolated maps for daily PM₁₀ concentration from May 1 to 9, 2005.

https://www.researchgate.net/publication/236011349_Spatio-temporal_analysis_and_interpolation_of_PM_10_measurements_in_Europe

Tools for reproducible research

1 A programmable **computing language**

1 e.g., the **R** project for statistical computing; **Python**

2 A **literate programming language** to mix code, output and text

■ e.g., **Markdown**², **knitr**³

3 A good **user interface** to use these

■ e.g., **R Studio**, **Jupyter Notebook**⁴ for Python

²<https://www.markdownguide.org>

³<https://yihui.name/knitr/>

⁴<https://jupyter.org>

Outline

- 1 Reproducible Research
- 2 The R Project
- 3 Some simple R
- 4 Literate Data Analysis
- 5 The R ecosystem
- 6 Final words

Why R?

- R is an **open-source** environment for **statistical computing, data manipulation and visualisation**;
- Statisticians have contributed over 15 000 specialised statistical procedures as **packages**;
- R and its packages are **freely-available over the internet**;
- R runs on Microsoft Windows, Unix[©] and derivatives Mac OS X and Linux;
- R is **fully programmable**, with modern computer language, **S**;
- Automation by user-written **scripts, functions** or **packages**;
- ...

Why R? (2)

- ...
- R has comprehensive **technical documentation**, user-contributed tutorials and textbooks showing **R code**;
- R syntax for model formulas are a standard, found in many documents;
- R can **import and export** in MS-Excel, text, fixed and delineated formats (e.g., CSV), databases (e.g., SQLite), vector and raster geographic coverages ...;
- R is fully supported in the R Markdown **literate programming environment**.

Some R packages for geography

`sp`, `sf` **structures** for spatial objects, spatial **overlay**,
subsetting, sampling . . .

`rgdal`, `proj4` **coordinate reference systems**, projections

`rgeos` **vector GIS** operations

`raster` manipulate spatial data in **raster format**

`gstat`, `fields`, `RandomFields`, `geoR` **model-based geostatistics**
(also Bayesian)

`spscosa`, `spsann` spatial **sampling**, simulated **annealing** (Brus)

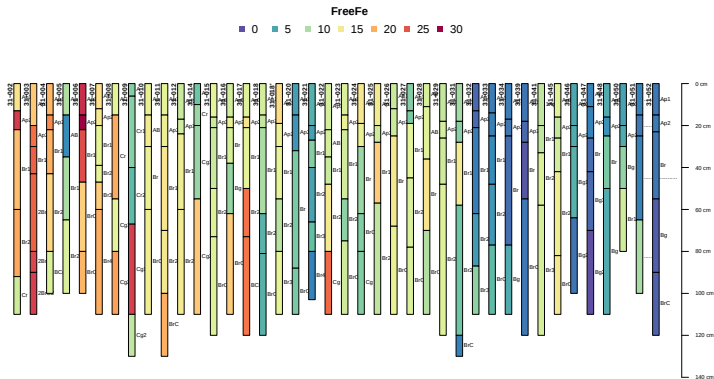
`spdep` spatial dependence on **lattice data** (like Geoda)

`spatstat` Spatial **Point Pattern Analysis**

`randomForest`, `ranger`, `Cubist`, `caret`, `nnet` . . . **machine learning**

R packages for pedology (soil science)

aqp Algorithms for Quantitative Pedology
soilDB access to USA soils databases



Outline

- 1 Reproducible Research
- 2 The R Project
- 3 Some simple R**
- 4 Literate Data Analysis
- 5 The R ecosystem
- 6 Final words

Expressions

R can be used as an **interactive calculator** at the `>` command prompt, with all the usual operators.

```
> 2*pi/360 # degrees to radians  
[1] 0.01745329
```

Any expression can be **saved** as an **object** in the **workspace** with the **assignment operator** `<-` and then re-used:

```
> deg2rad <- 2*pi/360  
> 30*deg2rad  
[1] 0.5235988
```

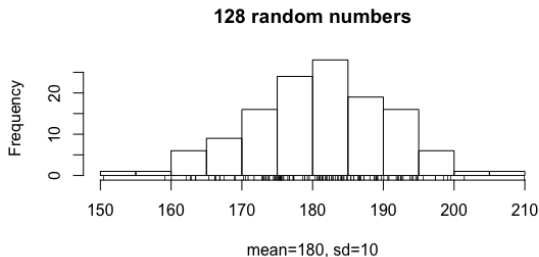
Functions

- Most computation in R depends on **functions**: **arguments** → **function** → **results**
- **Packages** all define functions, each function has some **help** (information on how to use) and **examples**
- Some simple functions: `mean`, `var`, `range`
- Some complex functions: `lm`, `glms`, `krige`,
- Some graphics functions: `plot`, `ggplot`, `hist`
- Some functions change their behaviour depending on the class of their argument, e.g. `summary`, `predict`

Examples of functions

```
> srs <- sample(rnorm(128, mean=180, sd=10), size=12, replace = FALSE)
> round(sort(srs), 1)
[1] 168.2 172.5 173.1 176.8 179.0 180.3 181.1 181.7 182.2 185.0 185.6 186.2

> hist(sample <- rnorm(128, 180, 10), breaks=16,
+       main="128 random numbers",
+       xlab="mean=180, sd=10")
> rug(sample)
```



Dataframes

A **matrix** with **named columns** (fields, attributes) and (optionally) named **rows** (cases, tuples).

```
> data(trees); class(trees) # a built-in example dataset
[1] "data.frame"
> dim(trees); colnames(trees)
[1] 31 3
[1] "Girth" "Height" "Volume"
> trees[1:3,] # access rows by row number
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
> trees[, 1] # access columns by column number
[1] 8.3 8.6 8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 11.3 11.4 11.4 11.7 12.
[17] 12.9 13.3 13.7 13.8 14.0 14.2 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0 20.
> trees[which.max(trees$Volume), ] # find largest tree, note $ operator
  Girth Height Volume
31 20.6     87    77
```

Vectorized computations

Almost never need to write **for** loops; most operations work over vectors **element-wise**.

```
> length(trees$Girth); length(trees$Height)
[1] 31
[1] 31
> summaryt(trees$Girth) # circumference inches, see metadata
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.30   11.05   12.90   13.25   15.25   20.60
> summary((trees$Girth*2.54)/pi) # convert to diameter cm
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 6.711   8.934  10.430  10.711  12.330  16.655
> summary(trees$Height/trees$Girth) # thinness/thickness index
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4.223   4.705   6.000   5.986   6.838   8.434
```

Model formulas

A compact way to describe the components of a model (linear, random forest, time series ...)

- ~ “depends on”
- + additive
- * interaction
- / nested
- remove a term

Examples:

```
log(Zn) ~ flood.frequency + distance*elevation
```

```
weight ~ grade/height
```

(these names must be defined in the data used for the model)

```

> data(trees)
> summary(trees)
      Girth      Height      Volume
Min.   : 8.30   Min.   :63   Min.   :10.20
1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
Median :12.90   Median :76   Median :24.20
Mean   :13.25   Mean   :76   Mean   :30.17
3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
Max.   :20.60   Max.   :87   Max.   :77.00
> summary(model <- lm(Volume ~ Girth*Height, data=trees))
Residuals:
      Min       1Q   Median       3Q      Max
-6.5821 -1.0673  0.3026  1.5641  4.6649
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.39632    23.83575   2.911  0.00713 **
Girth        -5.85585     1.92134  -3.048  0.00511 **
Height       -1.29708     0.30984  -4.186  0.00027 ***
Girth:Height  0.13465     0.02438   5.524 7.48e-06 ***
---
Residual standard error: 2.709 on 27 degrees of freedom
Multiple R-squared:  0.9756, Adjusted R-squared:  0.9728
F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16

```

Matrices

Various operators and functions are provided for **matrices** (similar to Matlab):

- `+`, `-`, `*`, `/` etc. work element-wise
- matrix multiplication: `%*%`, inner and outer vector products
- transposition: `t` function
- inversion: `solve` function
- spectral decomposition: `eigen` function
- Singular Value Decomposition: `svd` function
- principal components: `princomp`, `prcomp`

Random numbers

Support for random selection, **simulation**.

Density d , distribution function p , quantile function q , random generation r :

*norm

*unif

*binom

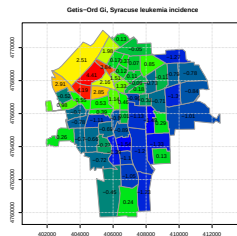
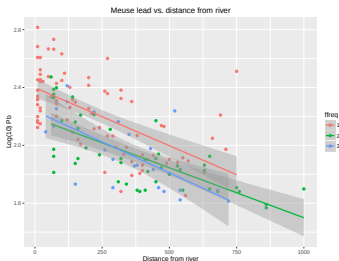
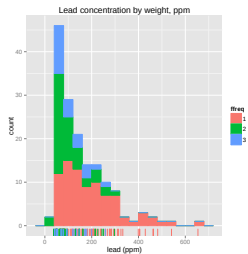
*beta

*chisq

*pois

...

Graphics



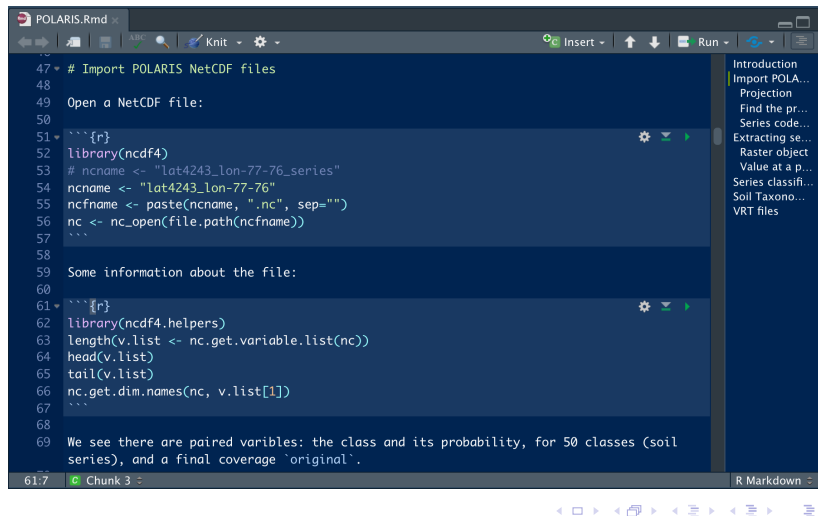
Outline

- 1 Reproducible Research
- 2 The R Project
- 3 Some simple R
- 4 Literate Data Analysis**
- 5 The R ecosystem
- 6 Final words

Literate Data Analysis

- Idea: mix **text** and **computations** in one document
- **text**: explain assumptions, choices made in computation, comments on results
 - can include computational results in text automatically
- **computation**: code that generates results, including text, tables, or graphics

Example – input in literate programming environment



The screenshot shows a literate programming environment with a dark blue background. The main editor displays R code for importing NetCDF files. The code is organized into sections separated by ````\n{r}` markers. The first section imports the `ncdf4` library and opens a NetCDF file. The second section imports `ncdf4.helpers` and displays information about the file. The third section is a text block explaining the data. The right sidebar shows a table of contents with the following items: Introduction, Import POLA..., Projection, Find the pr..., Series code..., Extracting se..., Raster object, Value at a p..., Series classifi..., Soil Taxono..., and VRT files. The bottom status bar indicates '61:7' and 'Chunk 3'.

```
47 # Import POLARIS NetCDF files
48
49 Open a NetCDF file:
50
51 ```\n{r}
52 library(ncdf4)
53 # ncname <- "lat4243_lon-77-76_series"
54 ncname <- "lat4243_lon-77-76"
55 ncfname <- paste(ncname, ".nc", sep="")
56 nc <- nc_open(file.path(ncfname))
57 ```
58
59 Some information about the file:
60
61 ```\n{r}
62 library(ncdf4.helpers)
63 length(v.list <- nc.get.variable.list(nc))
64 head(v.list)
65 tail(v.list)
66 nc.get.dim.names(nc, v.list[1])
67 ```
68
69 We see there are paired variables: the class and its probability, for 50 classes (soil
70 series), and a final coverage `original`.
```

61:7 Chunk 3 R Markdown

Example – compiled document

1 Introduction

2 Import POLARIS NetCDF files

2.1 Projection

2.2 Find the probable class

2.3 Series codes and names

3 Extracting series names at a point

4 Series classification

5 Soil Taxonomy for a set of points

6 VRT files

2 Import POLARIS NetCDF files

Open a NetCDF file:

```
library(ncdf4)
# ncname <- "lat4243_lon-77-76_series"
ncname <- "p_lat4243_lon-77-76"
ncfname <- paste(ncname, ".nc", sep="")
nc <- nc_open(file.path(ncfname))
```

Some information about the file:

```
library(ncdf4.helpers)
length(v.list <- nc.get.variable.list(nc))
```

```
## [1] 101
```

```
head(v.list)
```

```
## [1] "maxcl01_class" "maxcl01_prob" "maxcl02_class" "maxcl02_prob"
## [5] "maxcl03_class" "maxcl03_prob"
```

Example – input – with $\text{\LaTeX}$ formulas

```

666 # Sampling for mapping: K-means sampling using covariates
667
668 Now we use information from *covariates* that we think are correlated with the target variable. Our aim is to *create* strata based
on the values of the covariates. Presumably these combinations of covariates are better linked to the target than using just one
covariate. This is a form of *cluster analysis*. A popular method is k-means. This is explained in many references, e.g. Ch. 14 of:
669
670 Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning data mining, inference, and prediction
(2nd ed). New York: Springer. Retrieved from http://link.springer.com.proxy.library.cornell.edu/book/10.1007%2F978-0-387-84858-7
671
672 The concept is to find clusters that minimize the within-cluster variance and maximize the between-cluster variance. I.e., minimize:
673
674 
$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(j)=k} \|x_i - x_j\|^2 - \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2$$

675
676 This is minimized by assigning the  $N_k$  observations to the  $K$  clusters so that the average dissimilarity of the observations
within each cluster from that cluster's mean is minimized, over all clusters. Although this could be solved by exhaustive search,
that is not feasible for any reasonably-sized dataset, so there are various "greedy" algorithms, e.g., iterative descent from a
starting allocation.
677
678 ## Sample data
679
680 We use a study area from the Hunter Valley of New South Wales (AU), famous for its wines. The soil geography of this area has been
extensively studied by soil scientists from the University of Sydney. See for example:
681
682 Huang, J., McBratney, A. B., Malone, B. P., & Field, D. J. (2018). Mapping the transition from pre-European settlement to
contemporary soil conditions in the Lower Hunter Valley, Australia. Geoderma, 329, 27–42.
https://doi.org/10.1016/j.geoderma.2018.05.016
683
684 Load the file with discretisation of the study area and summarize it:
685
686

```

- 1 Design-based sampling for population statistics
- 2 Sampling to fit a linear regression
- 3 Sampling for mapping: grid sampling
- 4 Sampling for mapping: K-means sampling using covariates
- 4.1 Sample data
- 4.2 Construct clusters
- 4.3 Select sample
- 4.4 Visualizing the sampling plan in Google Earth.
- 5 Sampling for mapping: Model-based sampling for OK and KED
- 6 Sampling for mapping: Conditional Latin Hypercube sampling

4 Sampling for mapping: K-means sampling using covariates

Now we use information from *covariates* that we think are correlated with the target variable. Our aim is to *create* strata based on the values of the covariates. Presumably these combinations of covariates are better linked to the target than using just one covariate. This is a form of *cluster analysis*. A popular method is k-means. This is explained in many references, e.g. Ch. 14 of:

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning data mining, inference, and prediction (2nd ed). New York: Springer. Retrieved from <http://link.springer.com.proxy.library.cornell.edu/book/10.1007%2F978-0-387-84858-7>

The concept is to find clusters that minimize the within-cluster variance and maximize the between-cluster variance. I.e., minimize:

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(j)=k} \|x_i - x_j\|^2 - \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2$$

This is minimized by assigning the N observations to the K clusters so that the *average dissimilarity* of the observations within each cluster from that cluster's mean is minimized, over all clusters. Although this could be solved by exhaustive search, that is not feasible for any reasonably-sized dataset, so there are various “greedy” algorithms, e.g., iterative descent from a starting allocation.

4.1 Sample data

We use a study area from the Hunter Valley of New South Wales (AU), famous for its wines. The soil geography of this area has been extensively studied by soil scientists from the University of Sydney. See for example:

Huang, J., McBratney, A. B., Malone, B. P., & Field, D. J. (2018). Mapping the transition from pre-European settlement to contemporary soil conditions in the Lower Hunter Valley, Australia. *Geoderma*, 329, 27–42. <https://doi.org/10.1016/j.geoderma.2018.05.016>

Load the file with discretisation of the study area and summarize it:

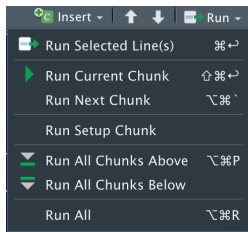
```
load("HunterValley4Practicals.RData")
# loaded object is `grdHunterValley`
summary(grdHunterValley)
```

#	Easting	Northing	elevation_m	slope_deg
##	##	##	##	##

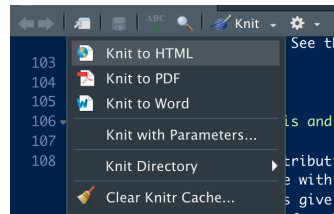
R Markdown

- A simple formatting language to **mix code and text**
- can be **compiled** to HTML, PDF, Word, websites ...
 - code is executed, output including graphics are written to the document, regular text is formatted
- code “chunks” are written inside a `{‘‘‘r} ... {‘‘‘}` block
- outside of these blocks are regular text
- computations can be included in the regular text with the `‘r ... ‘` syntax
- graphics commands automatically produce graphs in the output document

Running code chunks and compiling



Running code interactively



Compiling to a document

R Markdown example

(see demonstration)

Outline

- 1 Reproducible Research
- 2 The R Project
- 3 Some simple R
- 4 Literate Data Analysis
- 5 The R ecosystem**
- 6 Final words

R ecosystem

- the R project⁵ – manuals, FAQ, tutorials
 - CRAN⁶ – base R and packages
 - task views⁷ – compare packages for different applications
 - These are **mirrored** at many sites over the world.
- R Studio⁸ – environment to use R and R Markdown
- R Markdown⁹ – literate data analysis

⁵<https://www.r-project.org>

⁶<https://mirrors.tuna.tsinghua.edu.cn/CRAN/index.html>

⁷<https://mirrors.tuna.tsinghua.edu.cn/CRAN/web/views/>

⁸<https://www.rstudio.com>

⁹<https://rmarkdown.rstudio.com>

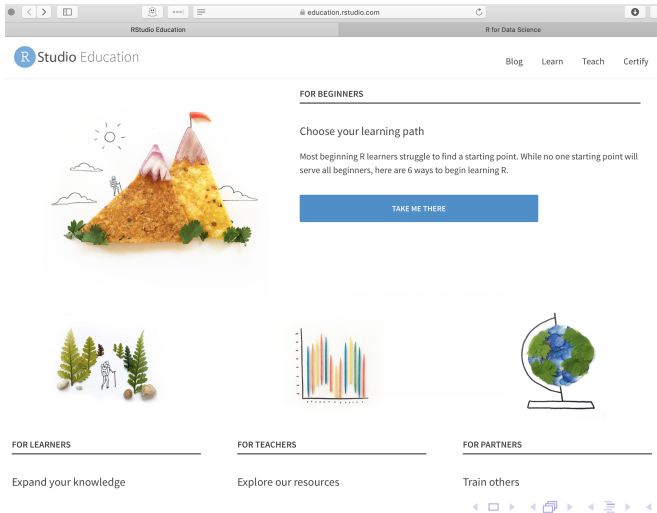
Learning R

- Introductions and tutorials
 - my R tutorials and applications¹⁰
- On-line help (within R and on the Internet)
- Contributed documentation
- Textbooks
- Task views
- R Journal, Mailing lists, user's conference

¹⁰[http:](http://www.css.cornell.edu/faculty/dgr2/pubs/list.html#pubs_in_R)

[//www.css.cornell.edu/faculty/dgr2/pubs/list.html#pubs_in_R](http://www.css.cornell.edu/faculty/dgr2/pubs/list.html#pubs_in_R)

Tutorials



The screenshot shows the RStudio Education website. At the top, there's a navigation bar with the RStudio logo and the word "Education". To the right of the logo, there are links for "Blog", "Learn", "Teach", and "Certify". Below the navigation bar, there's a large section titled "FOR BEGINNERS" with a sub-header "Choose your learning path". The text below this sub-header states: "Most beginning R learners struggle to find a starting point. While no one starting point will serve all beginners, here are 6 ways to begin learning R." Below this text is a blue button labeled "TAKE ME THERE". To the left of the "FOR BEGINNERS" section is a large illustration of a mountain made of food, with a flag on top, a sun, and clouds. Below this illustration is a smaller illustration of a person climbing a mountain. To the right of the "FOR BEGINNERS" section, there are three smaller sections: "FOR LEARNERS" with the sub-header "Expand your knowledge", "FOR TEACHERS" with the sub-header "Explore our resources", and "FOR PARTNERS" with the sub-header "Train others". Each of these sections has a corresponding illustration: a person climbing a mountain for learners, a bar chart for teachers, and a globe for partners. At the bottom of the page, there are navigation icons for back, forward, and search.

education.rstudio.com

RStudio Education

R for Data Science

Blog Learn Teach Certify

FOR BEGINNERS

Choose your learning path

Most beginning R learners struggle to find a starting point. While no one starting point will serve all beginners, here are 6 ways to begin learning R.

TAKE ME THERE

FOR LEARNERS

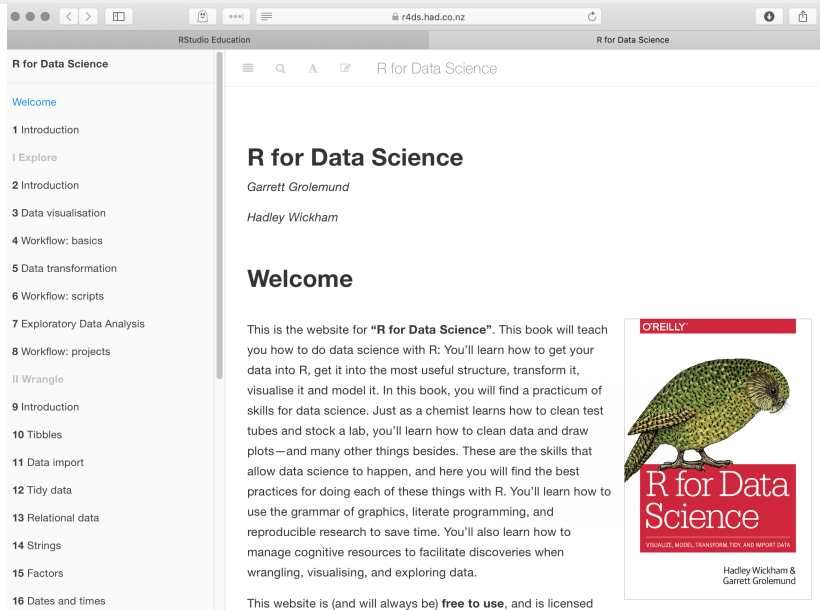
Expand your knowledge

FOR TEACHERS

Explore our resources

FOR PARTNERS

Train others



The screenshot shows a web browser window with the address bar displaying "r4ds.had.co.nz". The page title is "R for Data Science". The left sidebar contains a navigation menu with the following items: "Welcome", "1 Introduction", "I Explore", "2 Introduction", "3 Data visualisation", "4 Workflow: basics", "5 Data transformation", "6 Workflow: scripts", "7 Exploratory Data Analysis", "8 Workflow: projects", "II Wrangle", "9 Introduction", "10 Tibbles", "11 Data import", "12 Tidy data", "13 Relational data", "14 Strings", "15 Factors", and "16 Dates and times". The main content area has the heading "R for Data Science" by Garrett Grolemund and Hadley Wickham. Below this is a "Welcome" section with a paragraph of text. At the bottom right is a book cover for "R for Data Science" by Hadley Wickham & Garrett Grolemund, published by O'Reilly. The book cover features a green parrot and the text "VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA".

RStudio Education R for Data Science

R for Data Science

Welcome

1 Introduction

I Explore

2 Introduction

3 Data visualisation

4 Workflow: basics

5 Data transformation

6 Workflow: scripts

7 Exploratory Data Analysis

8 Workflow: projects

II Wrangle

9 Introduction

10 Tibbles

11 Data import

12 Tidy data

13 Relational data

14 Strings

15 Factors

16 Dates and times

R for Data Science

Garrett Grolemund

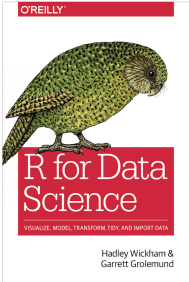
Hadley Wickham

Welcome

This is the website for “**R for Data Science**”. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

This website is (and will always be) **free to use**, and is licensed

O'REILLY



R for Data Science

VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA

Hadley Wickham & Garrett Grolemund

Reference sheets

R Markdown Cheat Sheet

learn more at rmarkdown.rstudio.com



.Rmd files

An R Markdown (.Rmd) file is a record of your research. It contains the code that a scientist needs to reproduce your work along with the narration that a reader needs to understand your work.

Reproducible Research

At the click of a button, or the type of a command, you can rerun the code in an R Markdown file to reproduce your work and export the results as a finished report.

Dynamic Documents

You can choose to export the finished report as a html/pdf, MS Word, ODT, RTF, or markdown document, or as a html or pdf based slide show.

Workflow

- 1 Open a new .Rmd file at File • New File • R Markdown. Use the wizard that opens to pre-populate the file with a template.
- 2 Write document by editing template.
- 3 Knit document to create report. Use knit button or `render()` to knit.
- 4 Preview Output in IDE window.
- 5 Publish (optional) to web or server. Synchronise publish button to accounts at:
 - rpubs.com
 - shinyapps.io
 - RStudio Connect
 Reload document. Find in document. File path to output document.
- 6 Examine build log in R Markdown console.
- 7 Use output file that is saved alongside .Rmd.

.Rmd structure

YAML Header

Optional section of render (e.g. pandoc) options written as key-value pairs (YAML).

- At start of file
- Between lines of ---

Text

Narration formatted with markdown, mixed with code. Each chunk:

- Begins with `'''(r)`
- ends with `'''`

Code chunks

Chunks of embedded code. Each chunk:

- Begins with `'''(r)`
- ends with `'''`

R Markdown will run the code and append the results to the doc.

It will use the location of the .Rmd file as the working directory.

Embed code with knitr syntax

Insert with `'''<code>'''`. Results appear as text without code.

One or more lines surrounded with `'''(r)<code>'''`. Place chunk options within curly braces, after `'''(r)<code>'''`.

Code chunks

```
'''(r echo=TRUE)
getVersion()'''
```

Global options

Set with knitr: `opts_chunk$set()` e.g.

```
opts_chunk$set(
  echo = TRUE,
  message = FALSE,
  results = 'markup',
  tidy = FALSE)

```

Inline code

```
knitr::opts_chunk$set(
  echo = TRUE,
  message = FALSE,
  results = 'markup',
  tidy = FALSE)
```

Cache - cache results for future knits (default = FALSE)

cache.path - directory to save cached results ("cache")

child - file(s) to knit and then include (default = NULL)

collapse - collapse all output into single block (default = FALSE)

dependencies - chunk dependencies for caching (default = NULL)

echo - Display code in output document (default = TRUE)

engine - code language used in chunk (default = 'R')

error - Display error messages in doc (TRUE) or stop render when errors occur (FALSE) (default = FALSE)

fig.align - 'left', 'right', or 'center' (default = 'default')

fig.cap - figure caption as character string (default = NULL)

fig.height, fig.width - Dimensions of plots in inches

highlight - highlight source code (default = TRUE)

message - display code messages in document (default = TRUE)

results - (default = 'markup') 'asis' - pass-through results 'hide' - do not display results 'hold' - put all results below all code

tidy - tidy code for display (default = FALSE)

Interactive Documents

Turn your report into an interactive Shiny document in 4 steps:

- 1 Add runtime: shiny to the YAML header.
- 2 Call Shiny input functions to embed input objects.
- 3 Call Shiny render functions to embed reactive output.
- 4 Render with markdown::run or click Run Document in RStudio IDE.

render()

Use markdown::render() to render/knit at cmd line. Important args:

Input - file to render

output, format - output, format

output_options - List of render options (as in YAML)

output_file - output file

output_dir - dir

params - list of params to use

envdir - environment to evaluate code chunks in

encoding - of input file

Embed a complete app into your document with shiny::shinyAppDir()

* Your report will rendered as a Shiny app, which means you must choose an html output format, like `html_document`, and serve it with an active R Session.

Textbooks

- Dalgaard, P. 2008. *Introductory Statistics with R*. Springer Verlag.
- Venables, W. N. & Ripley, B. D. 2002. *Modern applied statistics with S*. New York: Springer-Verlag, 4th edition¹¹
- Fox, J. 2011 *An R and S-PLUS Companion to Applied Regression*. Newbury Park: Sage.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An introduction to statistical learning: with applications in R*, Springer texts in statistics. Springer, New York.

¹¹<http://www.stats.ox.ac.uk/pub/MASS4/>

UseR! series

Springer is publishing a series¹² of **practical introductions with R code** to topics such as:

- data manipulation
- Bayesian analysis
- spatial data analysis
- time-series, space-time statistics
- interactive graphics
- machine learning
- applications, e.g., ecology, biomedicine, chemometrics, forestry ...

¹²<https://www.springer.com/series/6991?detailsPage=titles>



Applied Spatial Data Analysis with R

Authors ([view affiliations](#))

Roger S. Bivand, Edzer J. Pebesma, Virgilio Gómez-Rubio

Book

8

Citations

14

Mentions

1.7k

Readers

4

Reviews

125k

Downloads

Part of the [Use R!](#) book series (USE R)

Download book PDF



[Table of contents](#) (11 chapters)

[About this book](#)

Task views

Task Views are a summary by a task maintainer of which packages are best suited for certain tasks. Full list at <https://cran.r-project.org/web/views/index.html>. Examples:

- **Analysis of Spatial Data**¹³
- **Multivariate Statistics**¹⁴
- **Environmetrics** (Ecological & Environmental Data)¹⁵
- **Hydrology**¹⁶
- **Clustering**¹⁷

¹³<https://cran.r-project.org/web/views/Spatial.html>

¹⁴<https://cran.r-project.org/web/views/Multivariate.html>

¹⁵<https://cran.r-project.org/web/views/Environmetrics.html>

¹⁶<https://cran.r-project.org/web/views/Hydrology.html>

¹⁷<https://cran.r-project.org/web/views/Cluster.html>

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
Databases	Databases with R
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
ExtremeValue	Extreme Value Analysis
Finance	Empirical Finance
FunctionalData	Functional Data Analysis
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
Hydrology	Hydrological Data and Modeling
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
MissingData	Missing Data
ModelDeployment	Model Deployment with R
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
SpatioTemporal	Handling and Analyzing Spatio-Temporal Data
Survival	Survival Analysis
TeachingStatistics	Teaching Statistics
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

Outline

- 1 Reproducible Research
- 2 The R Project
- 3 Some simple R
- 4 Literate Data Analysis
- 5 The R ecosystem
- 6 Final words**

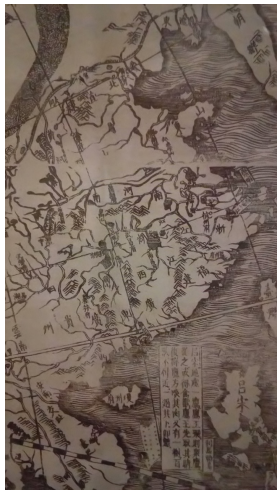
Never ever do these things!

- **Modify** an item in a primary database (e.g., Excel sheet) unless it is a clear data entry error, referring to the original field or lab paper sheets
 - modify (with explanation!) in the analysis
- **Remove** any items from a primary database
 - subset appropriately (with explanation!) in the analysis
- **Compute** any composite item in a primary database (e.g., sum of base cations)
 - compute the composite item in the analysis – the code shows exactly what was done
- **Transform, project, or resample** a GIS or raster coverage in an interactive GIS
 - use raster, rgdal, rgeos, sp, sf ... R packages

Do these things!

- Do all your data manipulation and analysis in a **reproducible research environment**;
- **Explain** the processing steps, choices and assumptions in text;
 - note that the code itself gives explanation of exactly how a result was obtained, but not assumptions or choices
- Produce graphics within the program – these change automatically if the processing changes

End



Amsterdam National (NL)
Maritime Museum
荷兰国家海洋博物馆
Verbiest world map 1674 坤輿
全图