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

Outline

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]

[1] Gentleman, 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 PM_{10} measurements in Europe

Task 1012 - Subtask 4

ETC/ACM Technical Paper 2011/10

Benedikt Gräler, Lydia Gerharz, Edzer Pebesma
ben.graeler@uni-muenster.de
November 2011

Including Erratum
(dd. 18 March 2013)

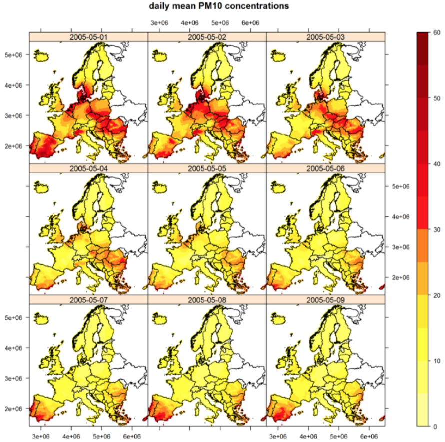


Figure 7: Interpolated maps for daily PM_{10} concentration from May 1 to 9, 2005.

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

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)

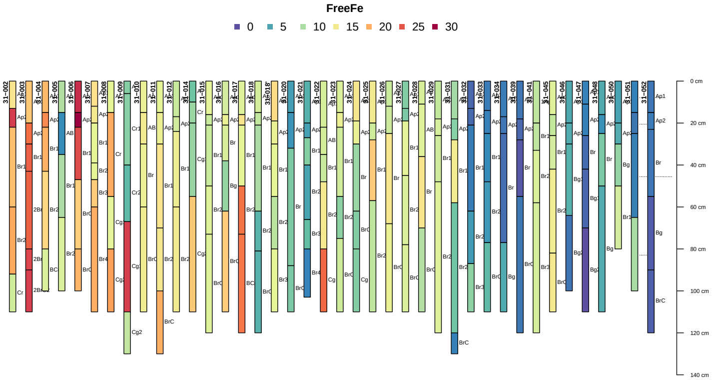
- ...
- R has comprehensive **technical documentation**, user-contributed tutorials and textbooks showing **R code**;
- R syntax for model formulas are a standard, found in many documents;
- R can **import and export** in MS-Excel, text, fixed and delineated formats (e.g., CSV), databases (e.g., SQLite), 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)
- spcosa, 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

Expressions

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

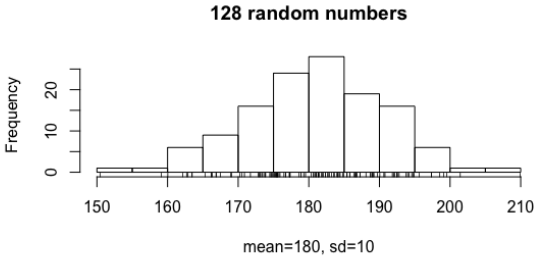
Functions

- 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`, `gls`, `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$Girth)  # circumfrence inches, see metadata
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.30   11.05   12.90   13.25   15.25   20.60
> summary((trees$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

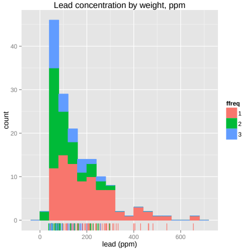
*beta

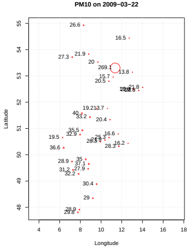
*chisq

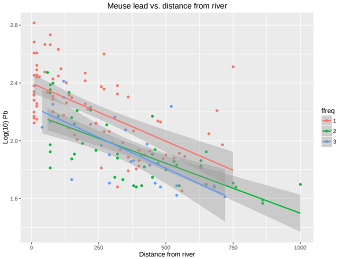
*pois

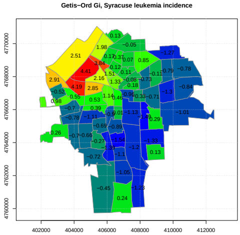
...

Graphics


Lead concentration by weight, ppm
count
lead (ppm)
ffreq
1
2
3


PM10 on 2009−03−22
Latitude
Longitude


Meuse lead vs. distance from river
Log(10) Pb
Distance from river
ffreq
1
2
3

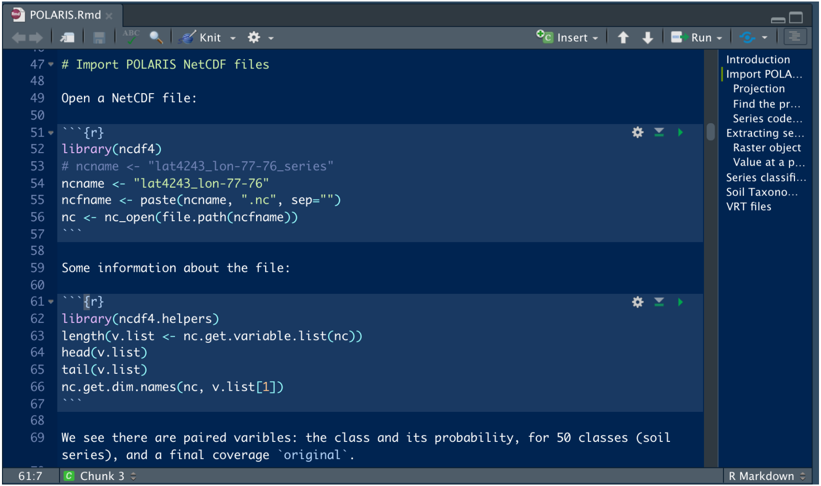

Getis−Ord Gi, Syracuse leukemia incidence

Outline

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



Example – compiled document

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:

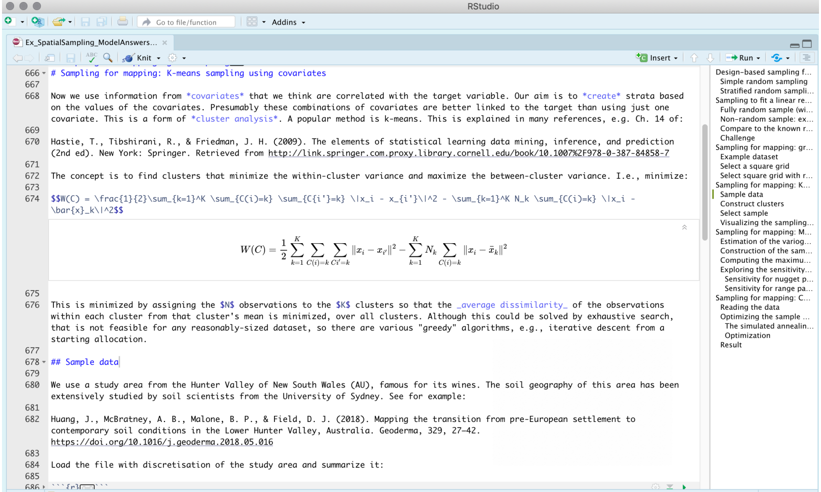
```
library(ncdf4.helpers)
length(v.list <- nc.get.variable.list(nc))
```

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

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

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

Example – input – with LaTeX formulas



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:

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

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:

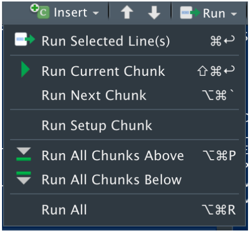
```
load("HunterValley4Practicals.RData")
# loaded object is `grdHunterValley`
summary(grdHunterValley)
```

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

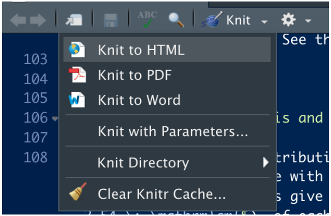
R Markdown

- 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

R ecosystem

- the R project[5] – manuals, FAQ, tutorials
  - CRAN[6] – base R and packages
  - task views[7] – compare packages for different applications
  - These are **mirrored** at many sites over the world.
- R Studio[8] – environment to use R and R Markdown
- R Markdown[9] – literate data analysis

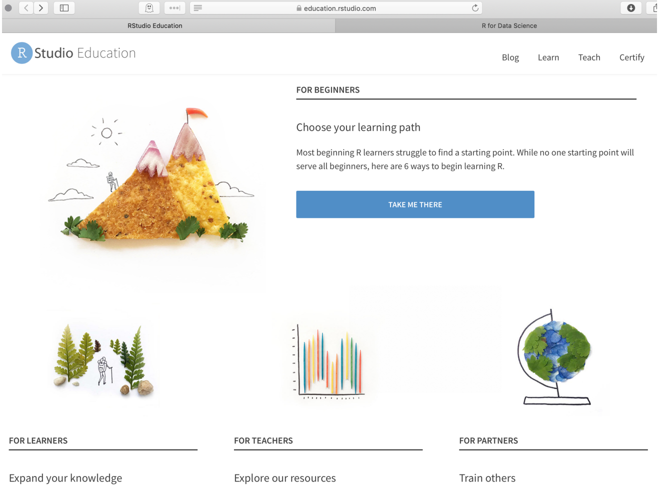
[5] https://www.r-project.org
[6] https://mirrors.tuna.tsinghua.edu.cn/CRAN/index.html
[7] https://mirrors.tuna.tsinghua.edu.cn/CRAN/web/views/
[8] https://www.rstudio.com
[9] https://rmarkdown.rstudio.com

Learning R

- Introductions and tutorials
  - my R tutorials and applications[10]
- On-line help (within R and on the Internet)
- Contributed documentation
- Textbooks
- Task views
- R Journal, Mailing lists, user's conference

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

Tutorials


education.rstudio.com
RStudio Education
R for Data Science
R Studio Education
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

r4ds.had.co.nz

RStudio Education | R for Data Science

R for Data Science

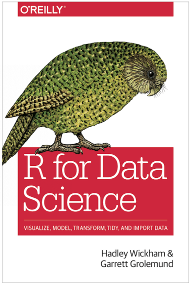
R for Data Science

R for Data Science

Garrett Grolemund

Hadley Wickham

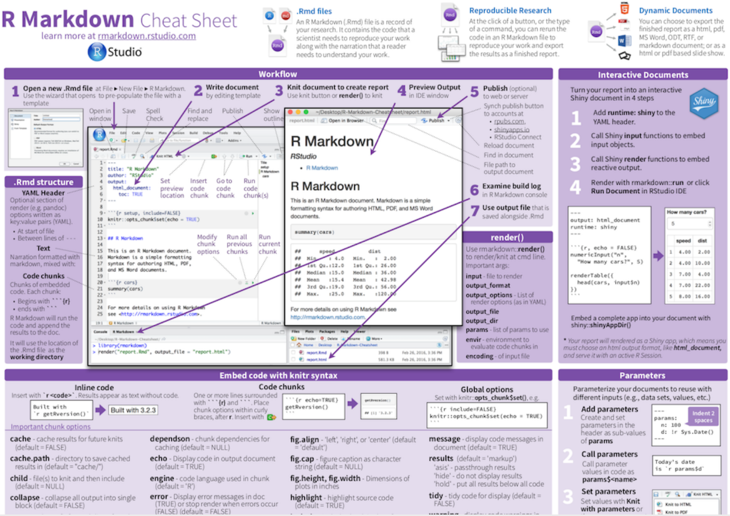
Welcome



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

Reference sheets



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 edition[11]
- 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/`

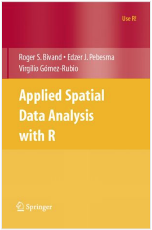
UseR! series

Springer is publishing a series[12] of **practical introductions with R code** to topics such as:

- data manipulation
- Bayesian analysis
- spatial data anlysis
- 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`

Springer Link



Applied Spatial Data Analysis with R

Authors (view affiliations)

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

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

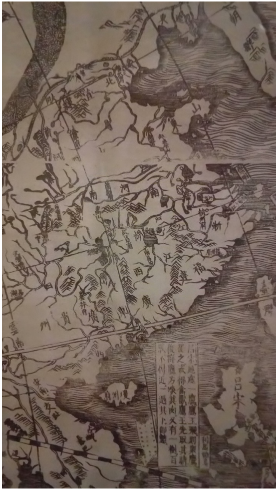
Never ever do these things!

- **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 and assumptions in text;
  - note that the code itself gives explanation of exactly how a result was obtained, but not assumptions or choiuces
- 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