R, an overview

Emanuel Huber

März 01, 2018

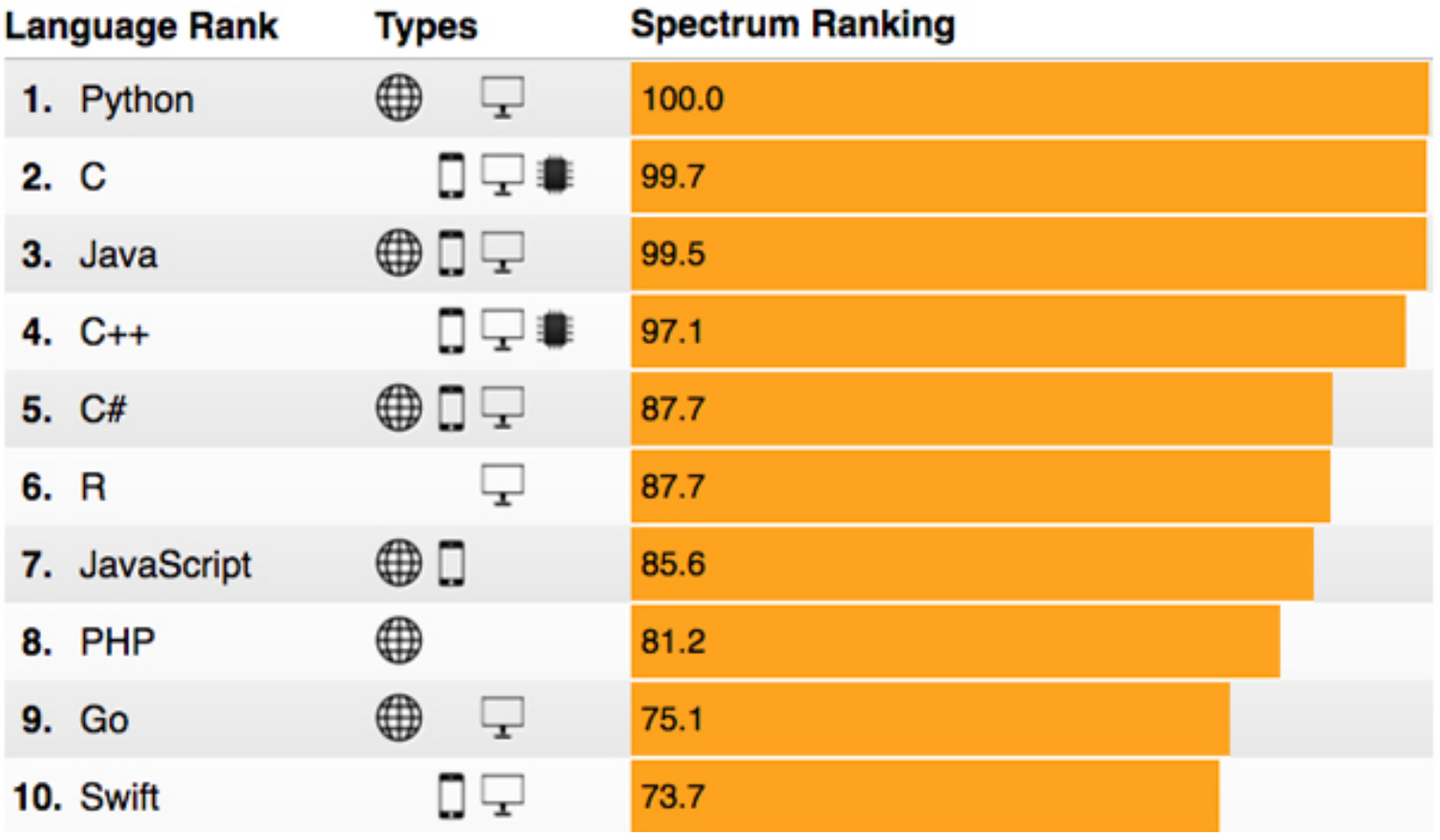
Outline

- Talk 1:
  1. Why R?
  2. R show case (incomplete)
- Talk 2: R language - some important concepts

Introduction

Programming language ranking by the IEEE (2017)

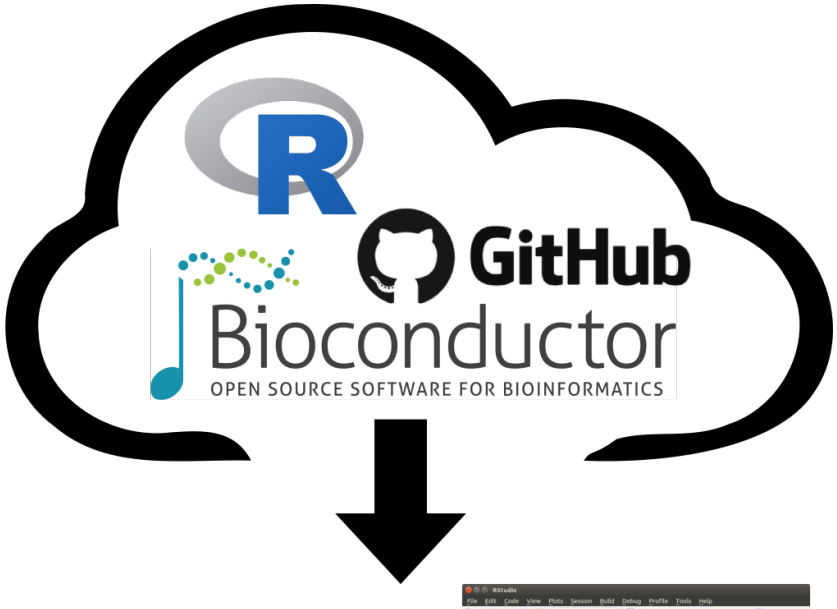
Institute of Electrical and Electronics Engineers ranking

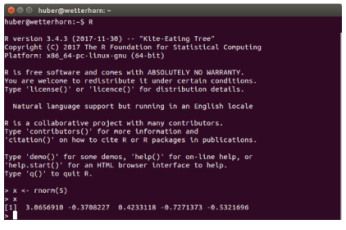


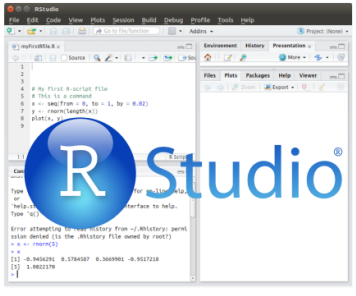
Why R is so popular?

- free and open-source (no licence)
- runs on Linux, Windows and MacOS.
- large community
- many packages available that are documented (help files + vignettes)
- can link C, C++, Fortran code
- excellent tools for data analysis (Google, Airbnb, Facebook, Microsoft...)
- high-quality graphics
- object-oriented programming

R Environment: GUI and packages







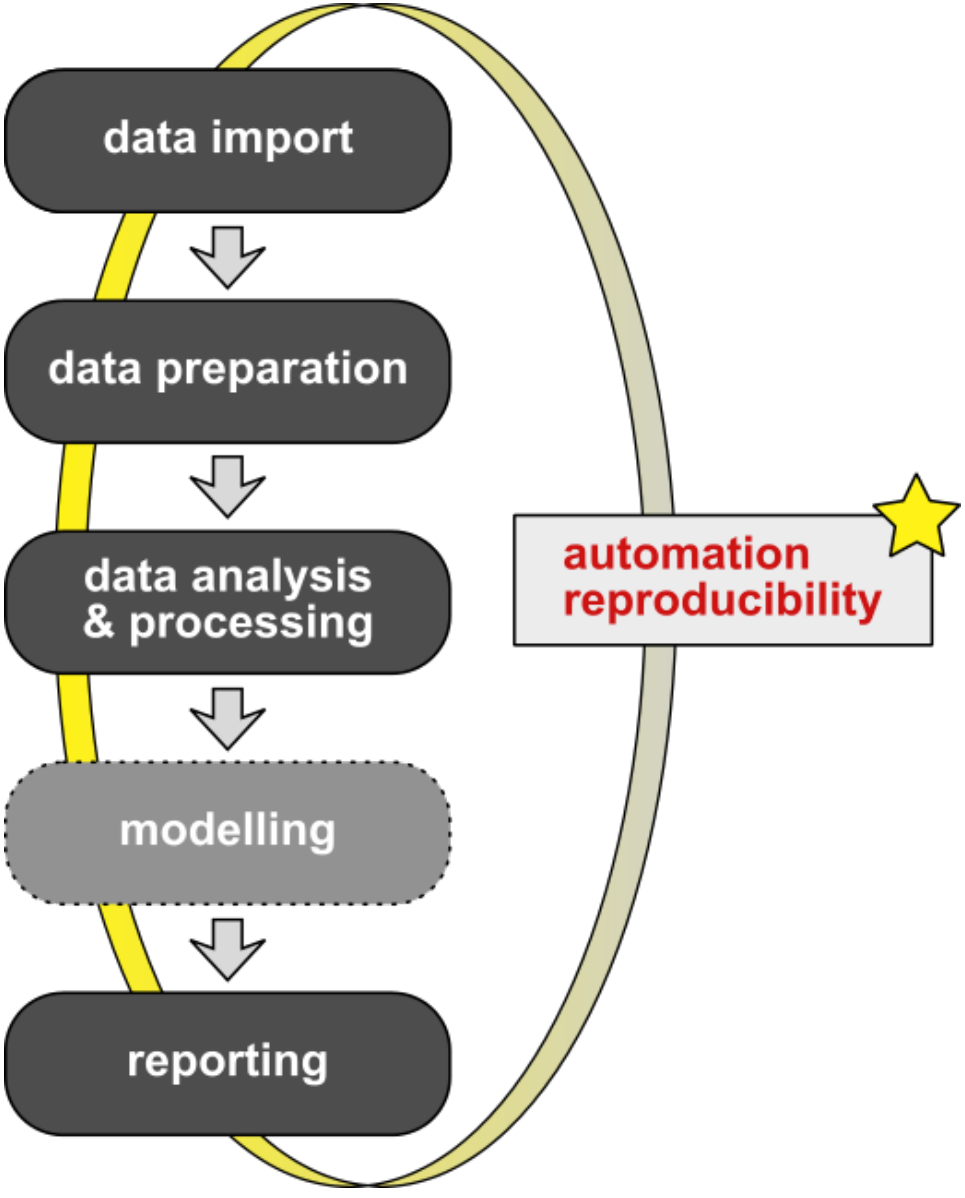
More than 12'000 packages!

CRAN Task Views

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
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
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
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
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

Potential use of R for AUG group

R strengths



- all-in-one tool
- no licence
- teaching, research, reporting
- automation and reproducible workflow

Data import/export

Resources - data import

Function	What It Does
`read.table()`	Reads any tabular data where the columns are separated `read.table(file = "filePath", sep = "\t", header = TRUE)`
`read.csv()`	A simplified version of read.table() to read CSV files. `read.csv(file = "filePath")`
`scan()`	Finer control over the read process when your data isn't tabular. `scan("filePath", skip = 1, nmax = 100)`
`readLines()`	Reads text from a text file one line at a time. `readLines("filePath")`
`read.fwf()`	Read a file with dates in fixed-width format. `read.fwf("filePath", widths = c(1, 2, 3)`
`readxl::read_excel()`	To read excel files (xls and xlsx), from `readxl` package `read_excel("filePath", sheet = "mtcars")`

Import/export

Data preparation

Resources - data cleaning

- **Package**
  - `dplyr`
- **Cheat sheet**
  - regular expression cheat sheet
  - `dplyr`: data transformation cheat sheet
- **Vignettes/Tutorials/book**
  - Tutorial: An introduction to data cleaning with R (53 p.)
  - Hands-On Data Science with R: Data Preparation
  - `dplyr` vignettes
  - Tutorial: Regular Expressions in R

Data cleaning (1)

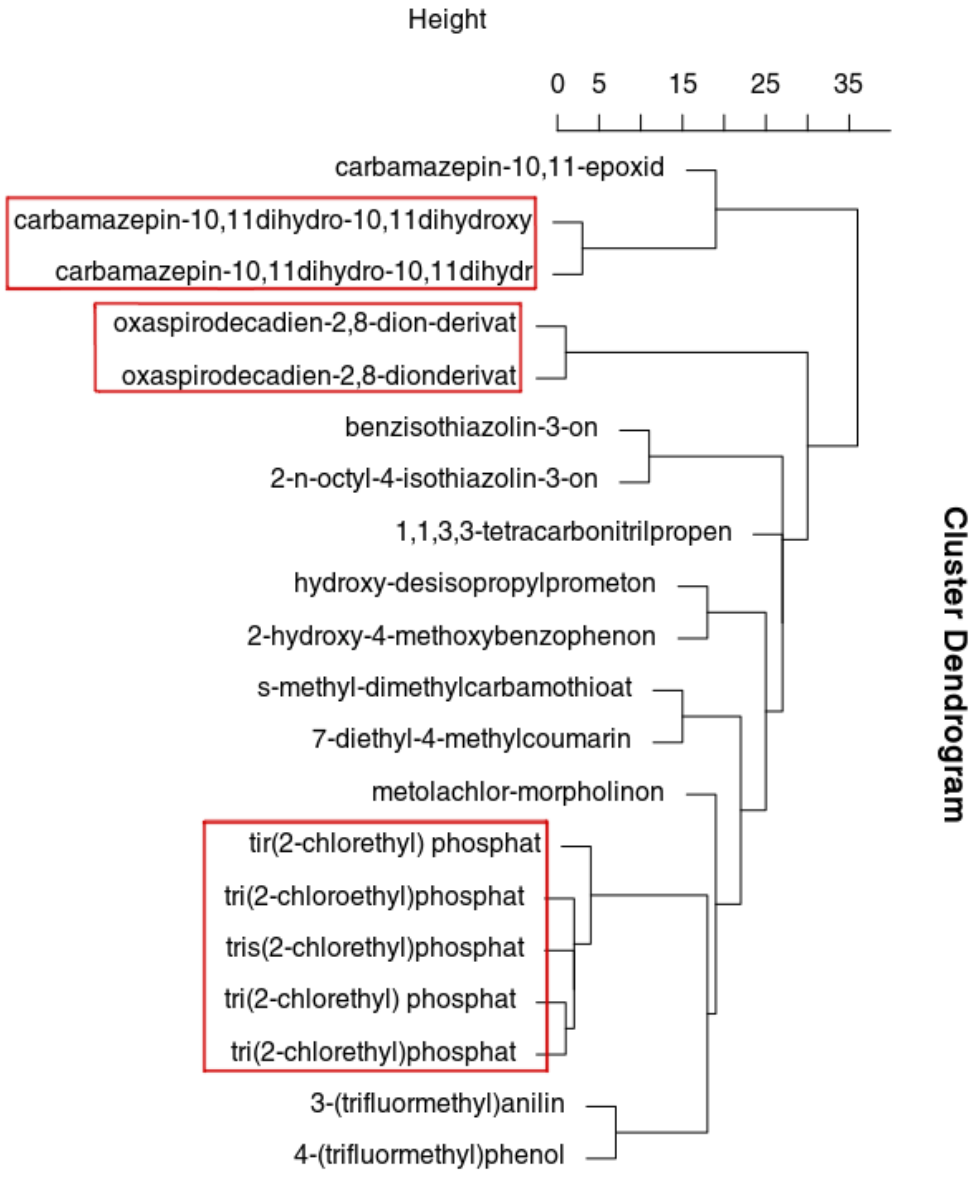
date	location	parameter	value
2011-01-12	RIVER	Head	2.31
2011-01-18	RIVER	nitrat	27.7
2011-01-18	GW_1	RIV_head	NA
2011-01-18	GW_1	nitrat	0.0083
2011-01-18	GW_1	sulfat	0
2011-01-18	GW_1	bromid	6.419
2011-01-18	GW_2	324/head	0.1226
2011-01-18	GW_2	nitrat	1.07
2011-01-18	GW_2	bromid	0
2011-03-02	RIVER	head	0.0194
2011-03-02	RIVER	nitrat	6.116
2011-03-02	RIVER	bromid	1.09
2011-03-02	GW_1	HEAD	0.0347

- remove rows/columns
- remove duplicates
- transform data
- correct for inconsistencies
- convert data type
- deal with missing values

- rename variables: `Head`, `RIV_head`, `324/head`, `head`, `HEAD` in `head`

```
pattern <- "^RIV_|^[[:digit:]]+/"
x$parameter <- sub(pattern, "", x$parameter)
x$parameter <- tolower(x$parameter)
```

Data cleaning (2)



detect typo errors

- distance between strings
- dendrogram visualisation
- partial string matching

Resources - Data shaping

- **Package**
  - `tidyr`
- **Cheat sheet**
  - `tidyr`: Tidy Data (see p. 2)
- **Tutorials/book**
  - Concepts of data tidying are well explained in the chapter Data Tidying from the book Data science with R.

Data shaping

"Messy" data

date	location	parameter	value
2011-01-12	RIVER	head	2.31
2011-01-18	RIVER	nitrat	27.7
2011-01-18	GW_1	head	NA
2011-01-18	GW_1	nitrat	0.0083
2011-01-18	GW_1	sulfat	0
2011-01-18	GW_1	bromid	6.419
2011-01-18	GW_2	head	0.1226
2011-01-18	GW_2	nitrat	1.07
2011-01-18	GW_2	bromid	0
2011-03-02	RIVER	head	0.0194
2011-03-02	RIVER	nitrat	6.116
2011-03-02	RIVER	bromid	1.09
2011-03-02	GW_1	head	0.0347

Tidy data satisfies three rules (see Data Science with R):

- Each variable in the data set is placed in its own column
- Each observation is placed in its own row
- Each value is placed in its own cell

Tidy data

date	location	head	nitrat	sulfat	bromid
2011-01-12	RIVER	2.31	27.7	NA	NA
2011-01-18	GW_1	NA	0.0083	0	6.419
2011-01-18	GW_2	0.1226	1.07	NA	0
2011-03-02	RIVER	0.0194	6.116	NA	1.09
2011-03-02	GW_1	0.0347	NA	NA	NA

```
x_tidy <- tidyr::spread(x, parameter, value)
```

reverse process possible with `tidyr::gather()`

Data analysis and processing

Data analysis

- data analysis
  - statistics
  - linear regression
  - boxplot
  - pairs-plot
  - correlation plot
  - dimensionality reduction (PCA, MDS, CCA, etc.)
  - cluster analysis

Time series - resources

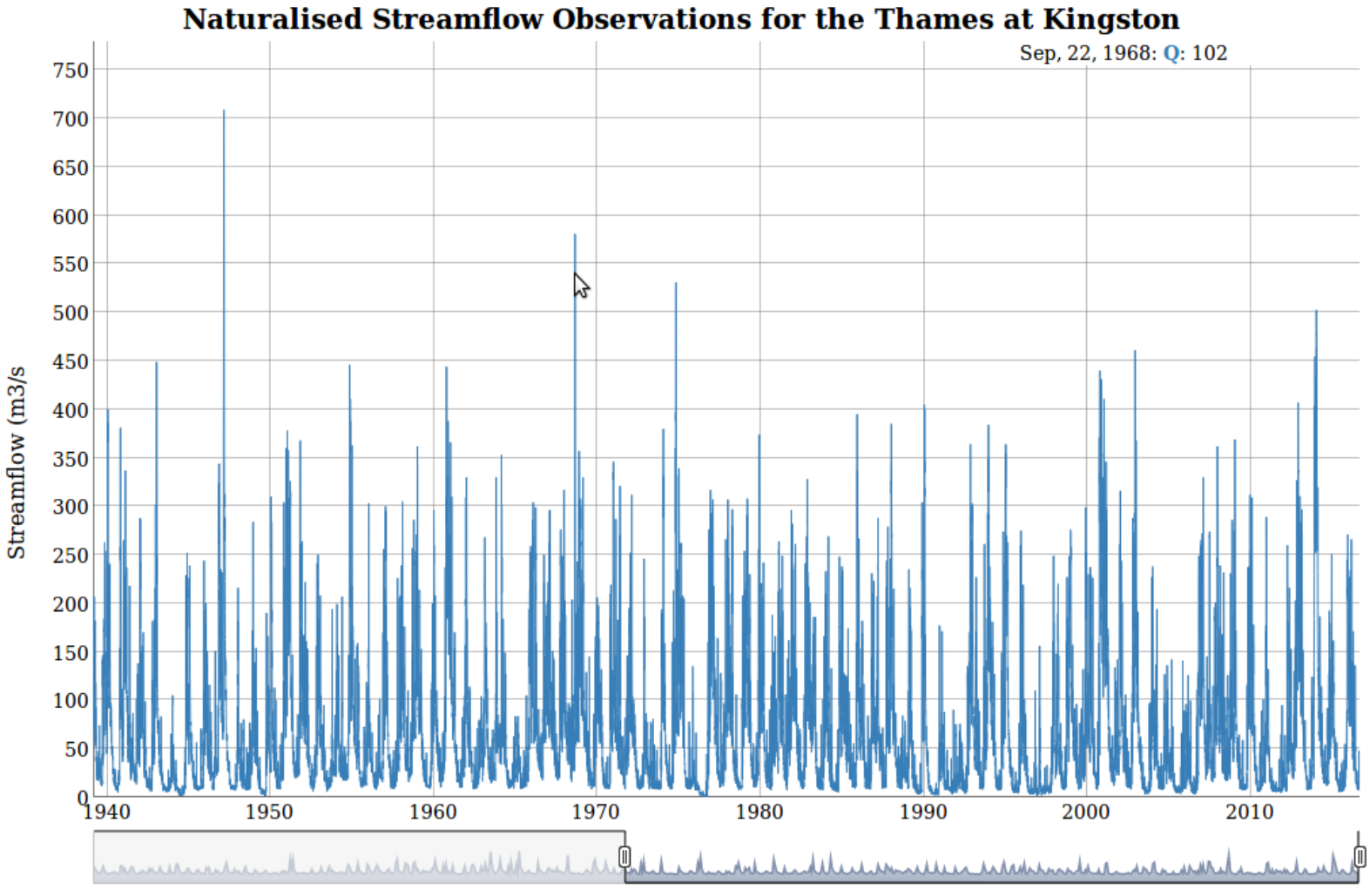
- **Package**
  - "eXtensible Time Series" `xts` that extends the `zoo` package
  - `zoo` for regular and irregular Time Series
  - `lubridate` to deal with date and time
- **Cheat sheet**
  - eXtensible Time Series: `xts`
  - How to deal with date and time: `lubridate` cheat sheet
- **Vignette/tutorials/book**
  - `xts` vignettes
  - `zoo` vignettes
  - Dates and Times Made Easy with `lubridate`

See also

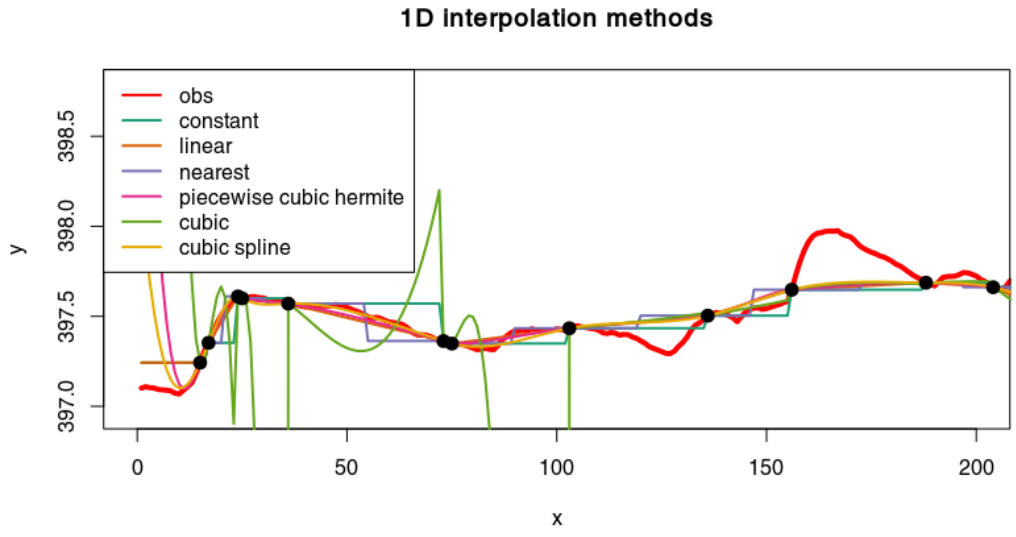
- [pracma]((https://cran.r-project.org/web/packages/pracma/index.html) for
- [signal]((https://cran.r-project.org/web/packages/signal/index.html) for
- check
- Filling in missing values?

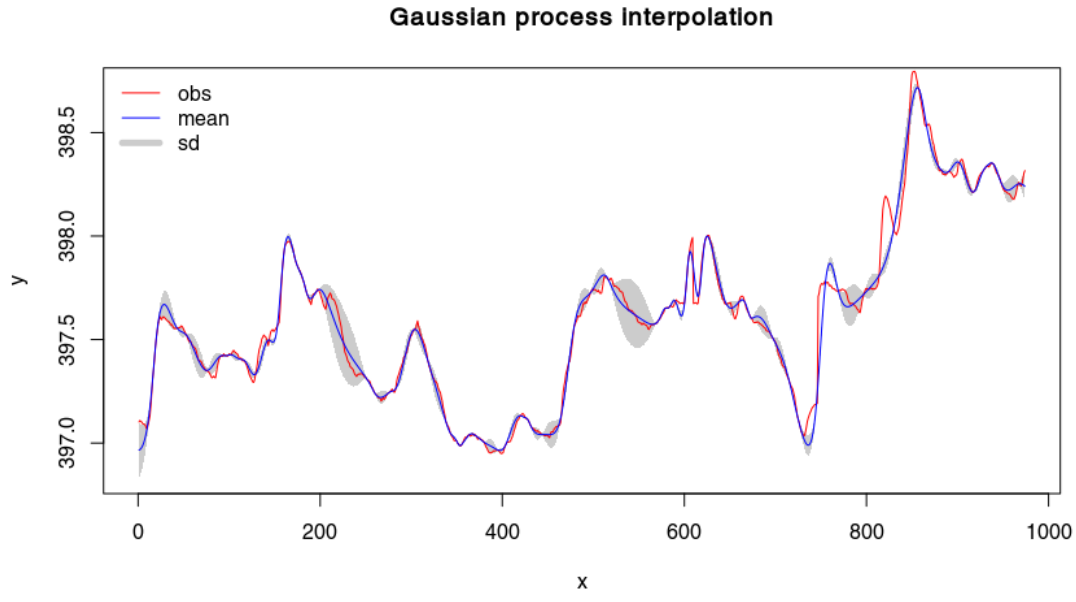
Time series - Dynamic visualisation

Dynamic time series visualisation with package `dygraph`

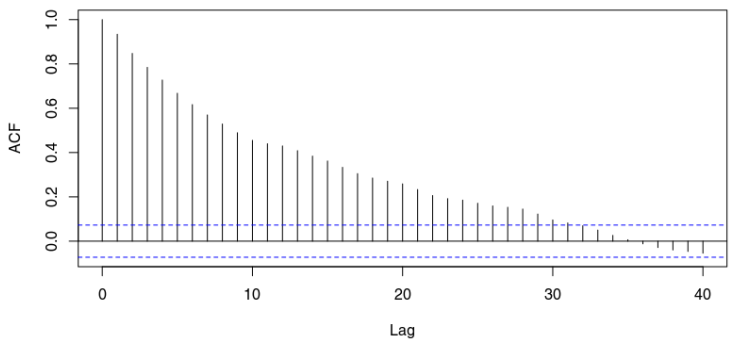


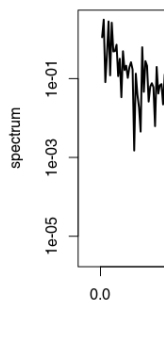
Time series - Interpolation

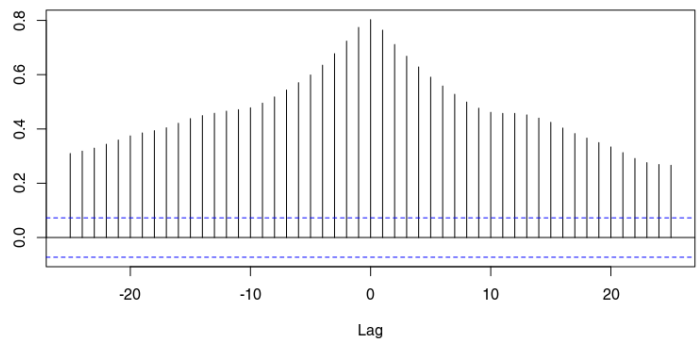


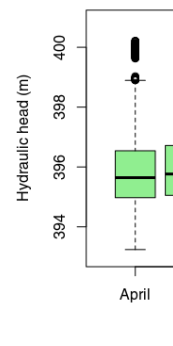


Time series - statistics



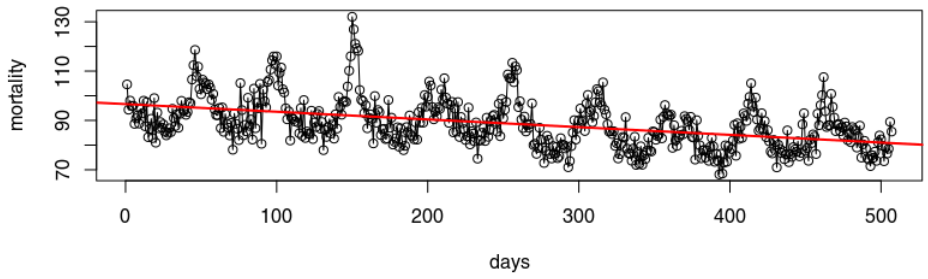




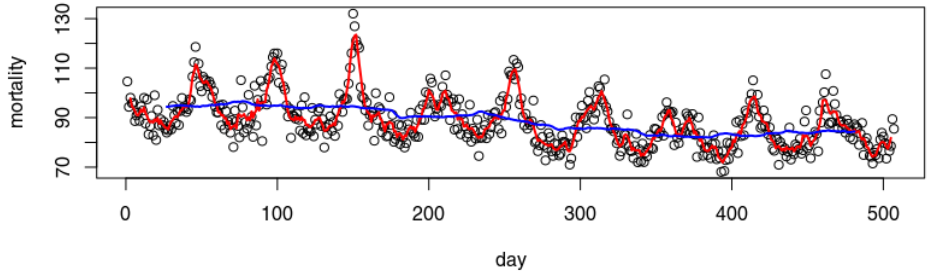


Time series - Regression and smoothing

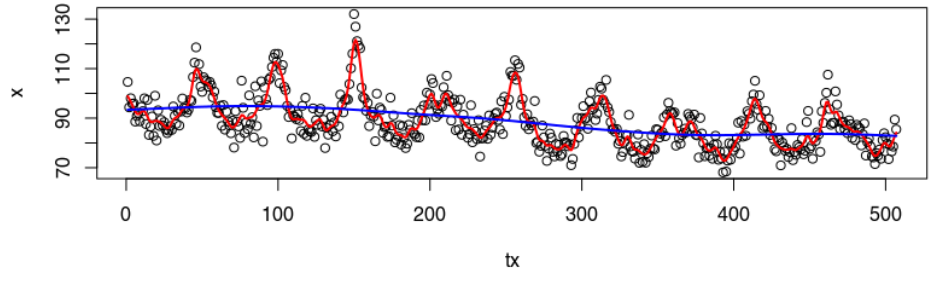
Regression



Moving average

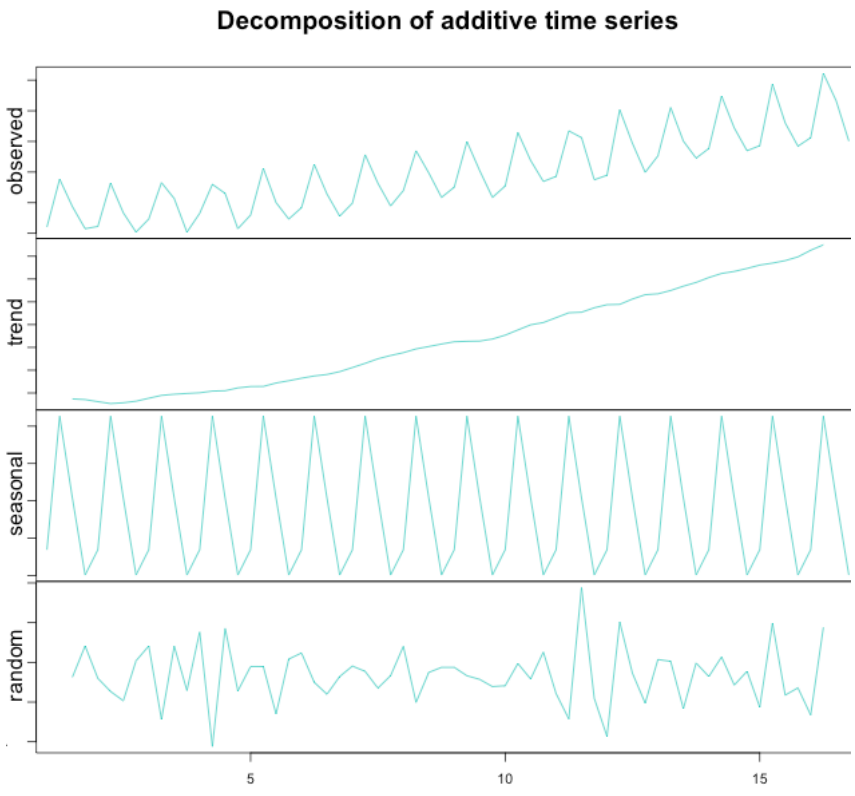


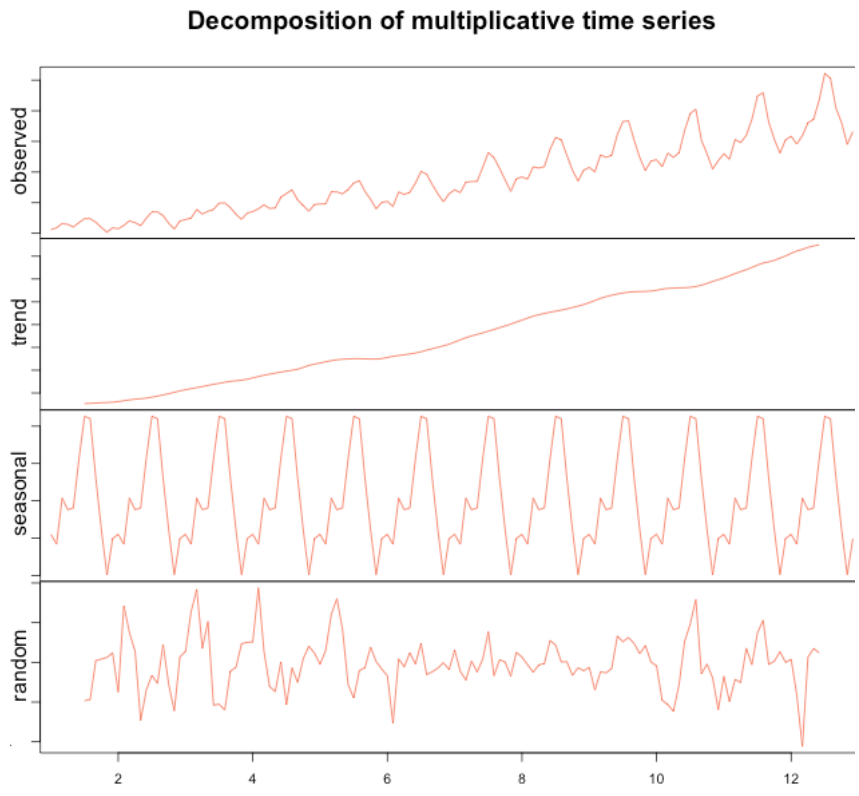
Smoothing



source: book "Time Series Analysis and Its Applications with R examples"

Time series - Trend and seasonal components

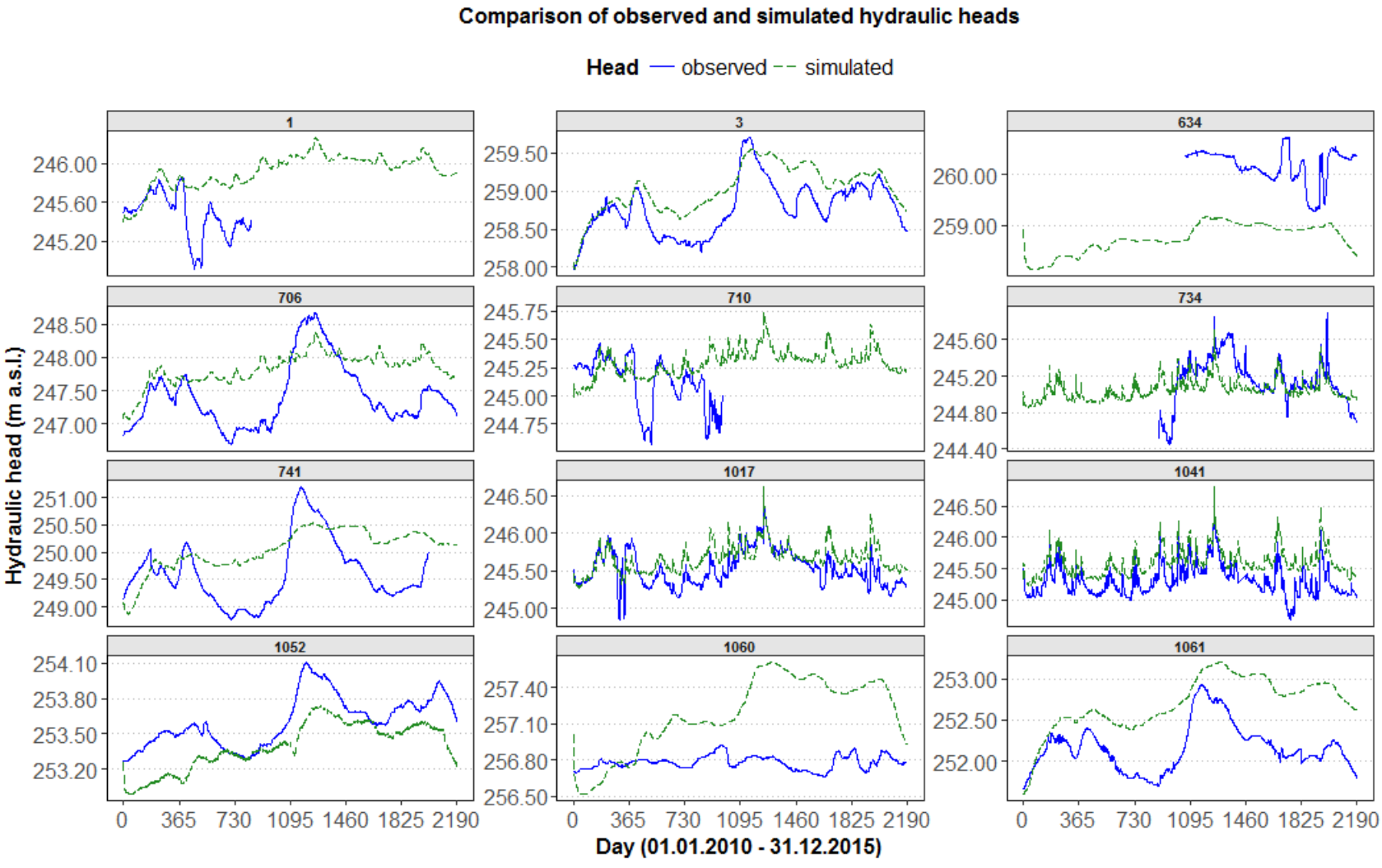





```
x_dcp_add <- decompose(x, "additive")
x_dcp_mul <- decompose(x, "multiplicative")
```

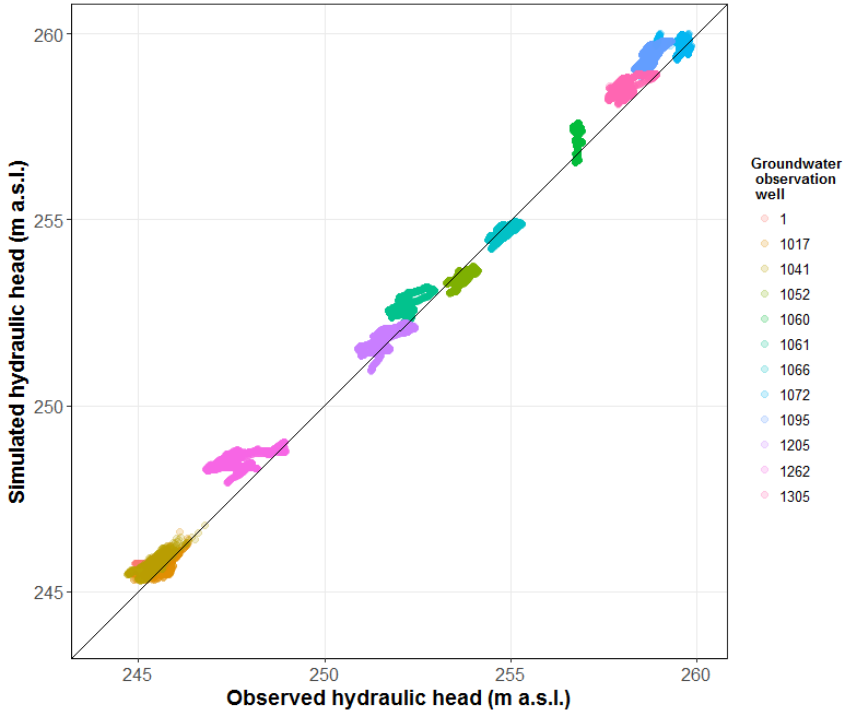
online tutorial

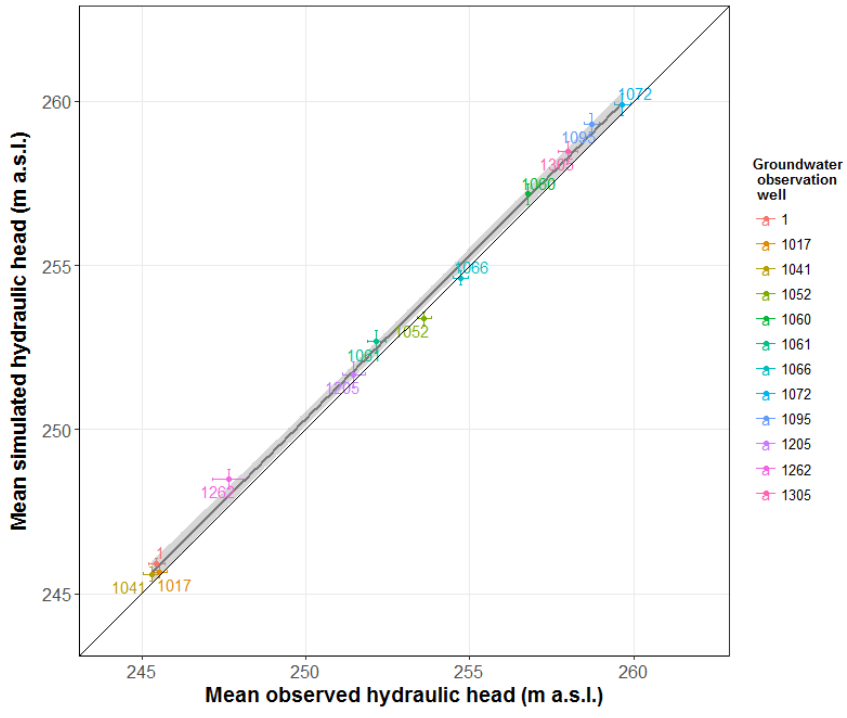
Time series - Import FEFLOW data



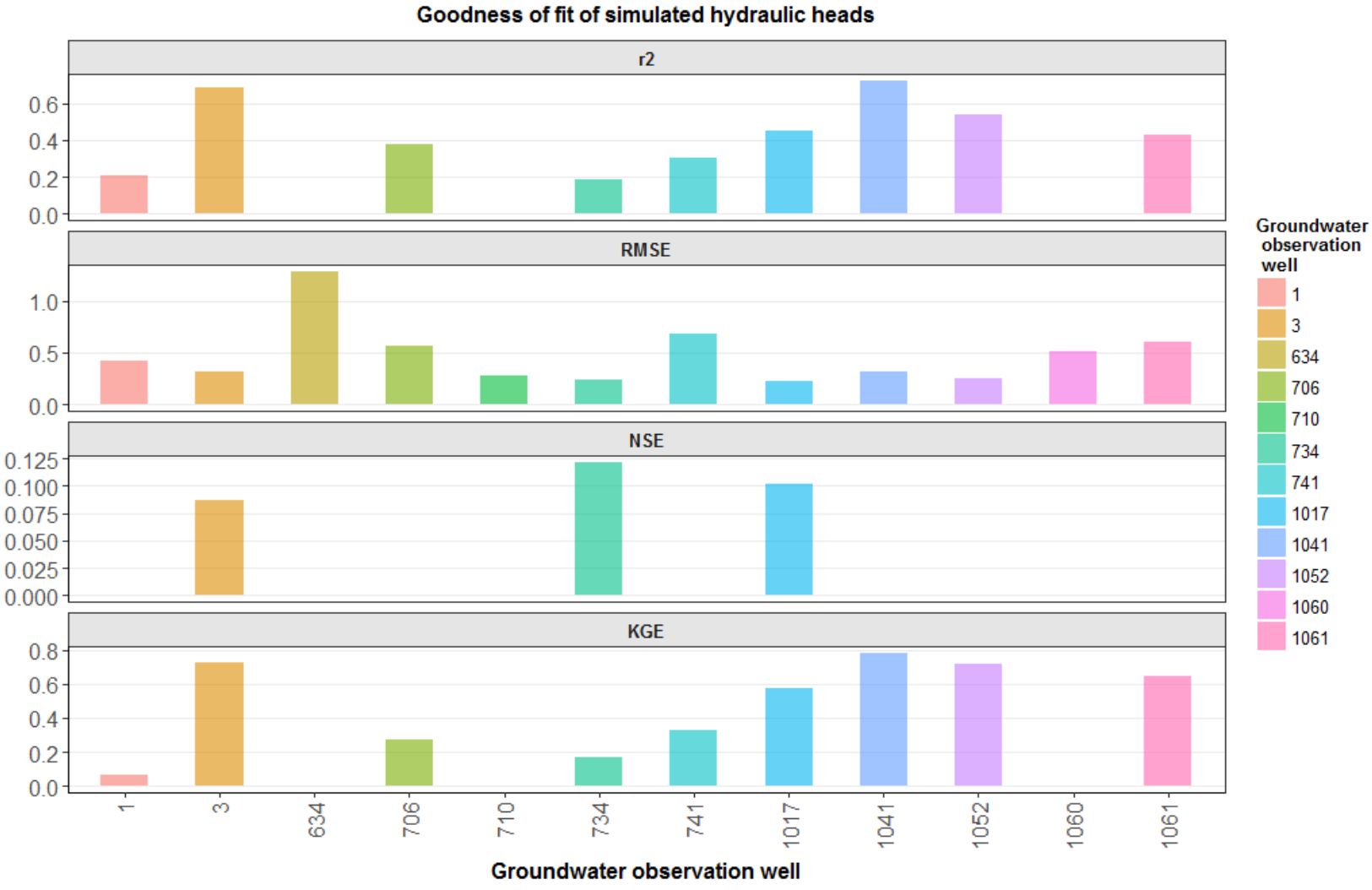
source: MM

Time series - FEFLOW: Simulated vs observed

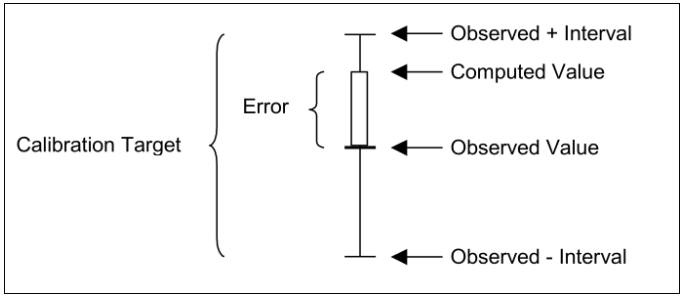


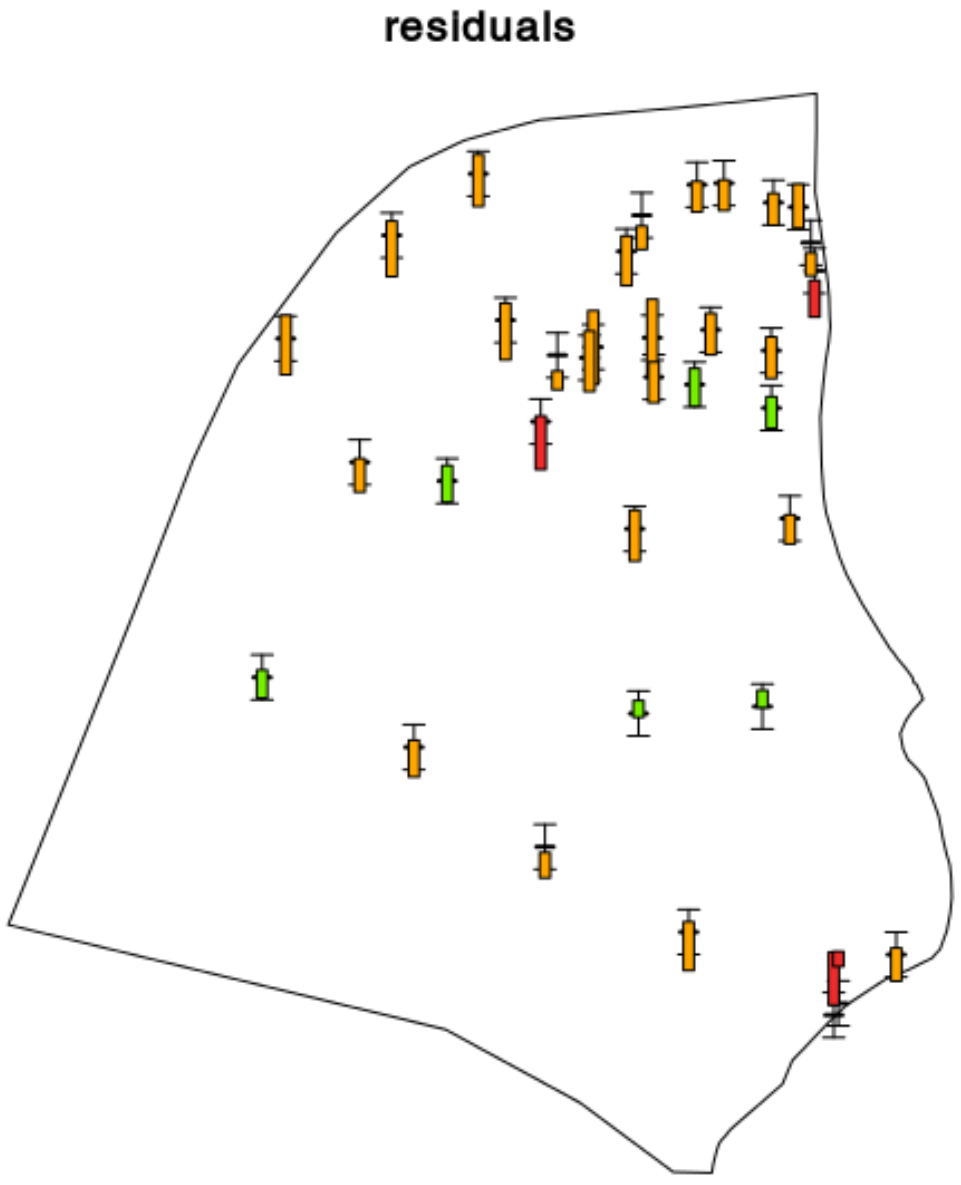


Time series - FEFLOW: Goodness-of-fit

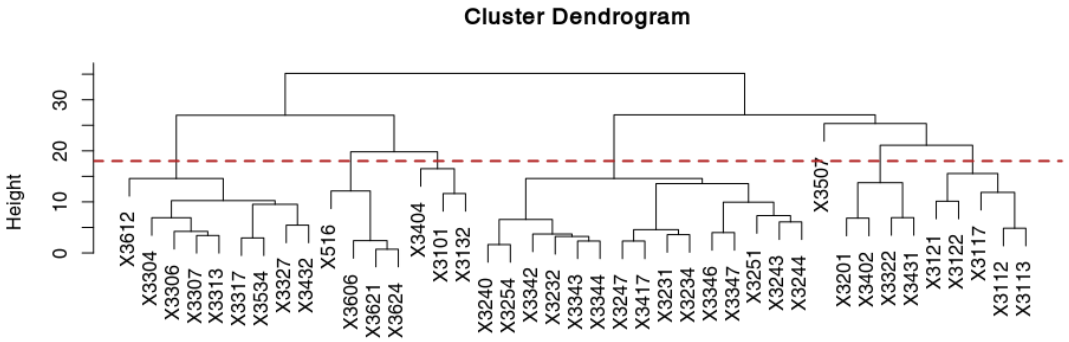


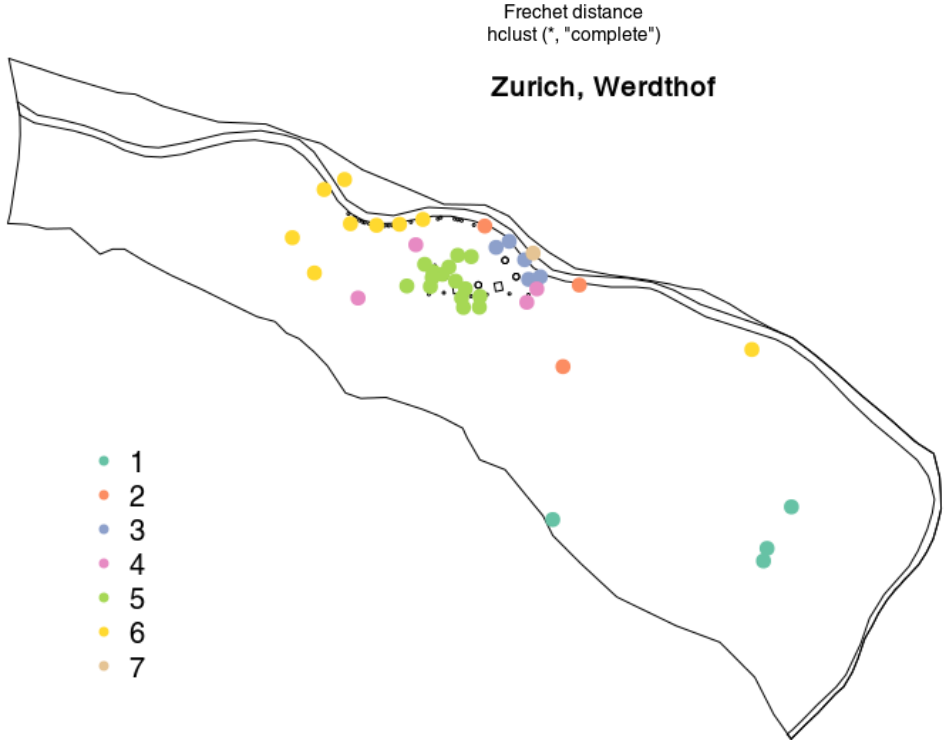
Time series - FEFLOW: Residual visualisation *à la GMS*





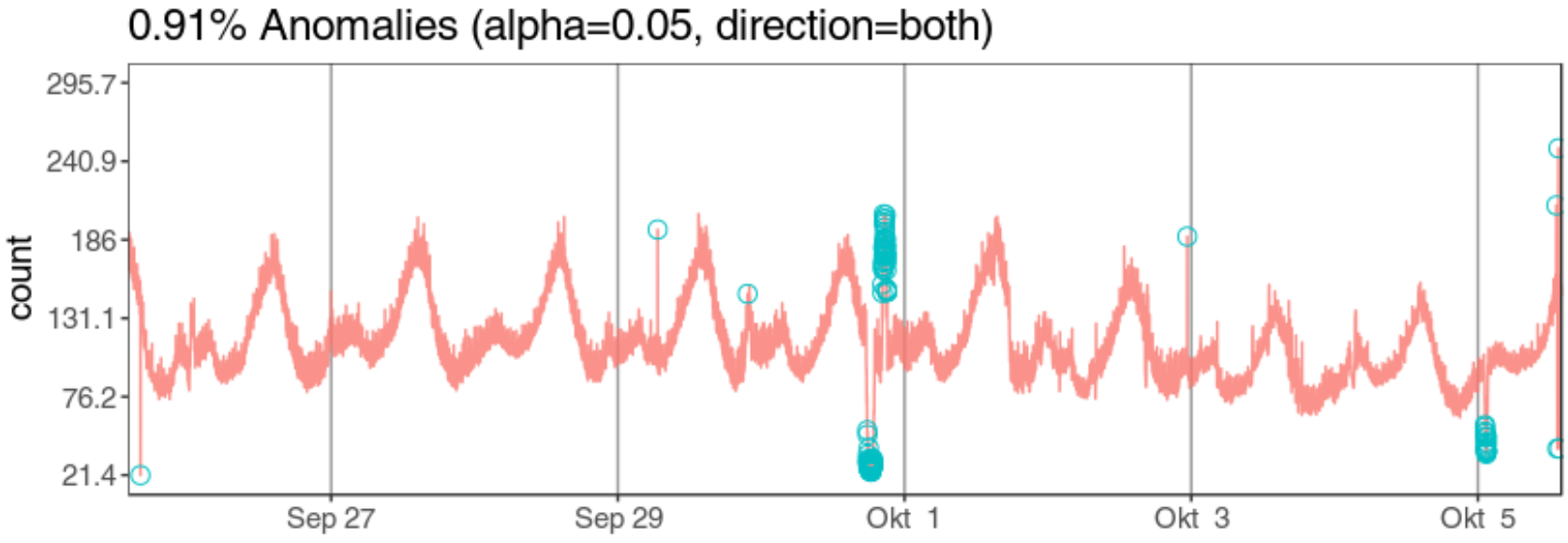
Time series - Clustering





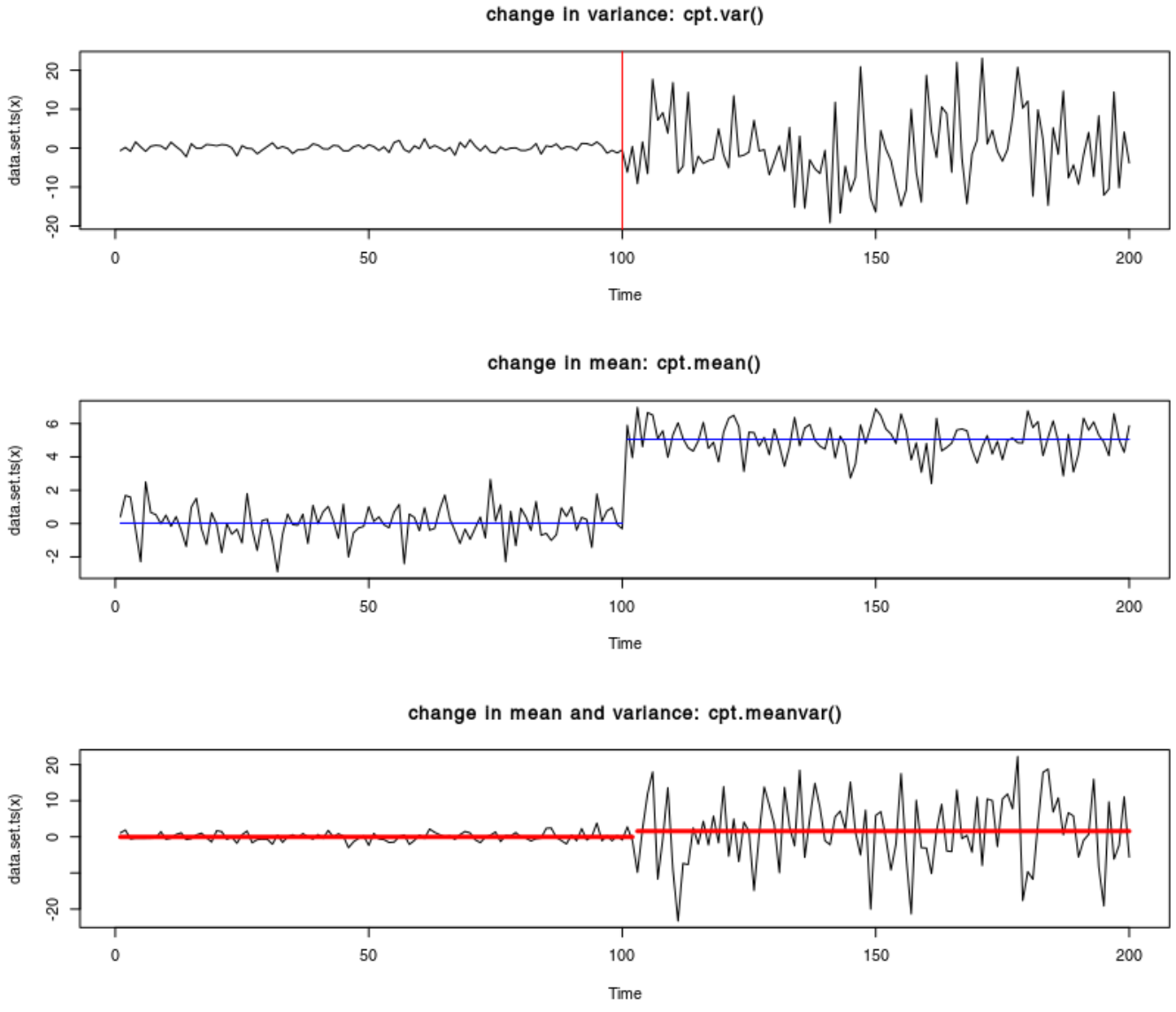
- remove obs with too many `NA`'s
- interpolate missing data `NA`
- compute distances between obs
- cut dendrogram –> classes
- plot onto map

Time series - Anomaly detection



source

Time series - change detection



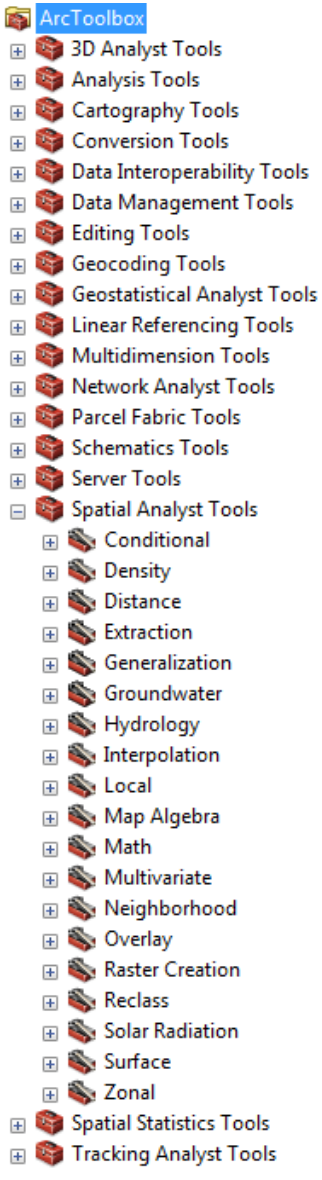
package `changepoint`

R GIS - resources

R is a GIS software (sp, rgeos, raster, rgdal) see also `sf` : chjeck

- **Package**
  - `sf` package for spatial vector data. Binds to GDAL for reading and writing data, to GEOS for geometrical operations, and to Proj.4 for projection conversions and datum (*will remplace sp/rgeos/rgdal in the long term* link)
  - `raster` package for raster, multi-band raster, ...
  - lidR for airborne LiDAR data manipulation and visualization for forestry application
  - packages to interact with existing GIS software link
    - `spgrass6`: Provides an interface between R and GRASS 6+. Allows for running R from within GRASS as well as running GRASS from within R.
    - `rgrass7`: Same as spgrass6, but for the latest version of GRASS, GRASS 7.
    - `RPyGeo`: A wrapper for accessing ArcGIS from R. Utilizes intermediate python scripts to fire up ArcGIS. Hasn't been updated in some time.
    - `RSAGA`: R interface to the command line version of SAGA GIS.
- **Cheat sheet**
  - nothing...
- **Tutorials/book**
  - Book: geocomputation with R
  - Spatial Data Analysis and Modeling with R

R GIS



Tool

- manipulation raster
  - raster interpolations
  - multi-band raster
  - slope, flow direction, hillshade
  - ...
- manipulation feature
  - attribute manipulations
  - spatial operations
  - geometric operations

Still missing

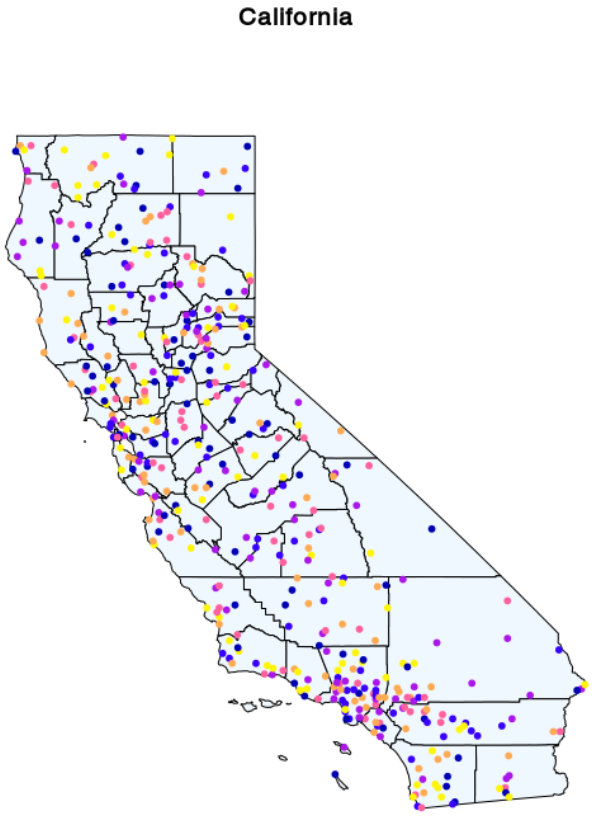
- TIN
- watershed

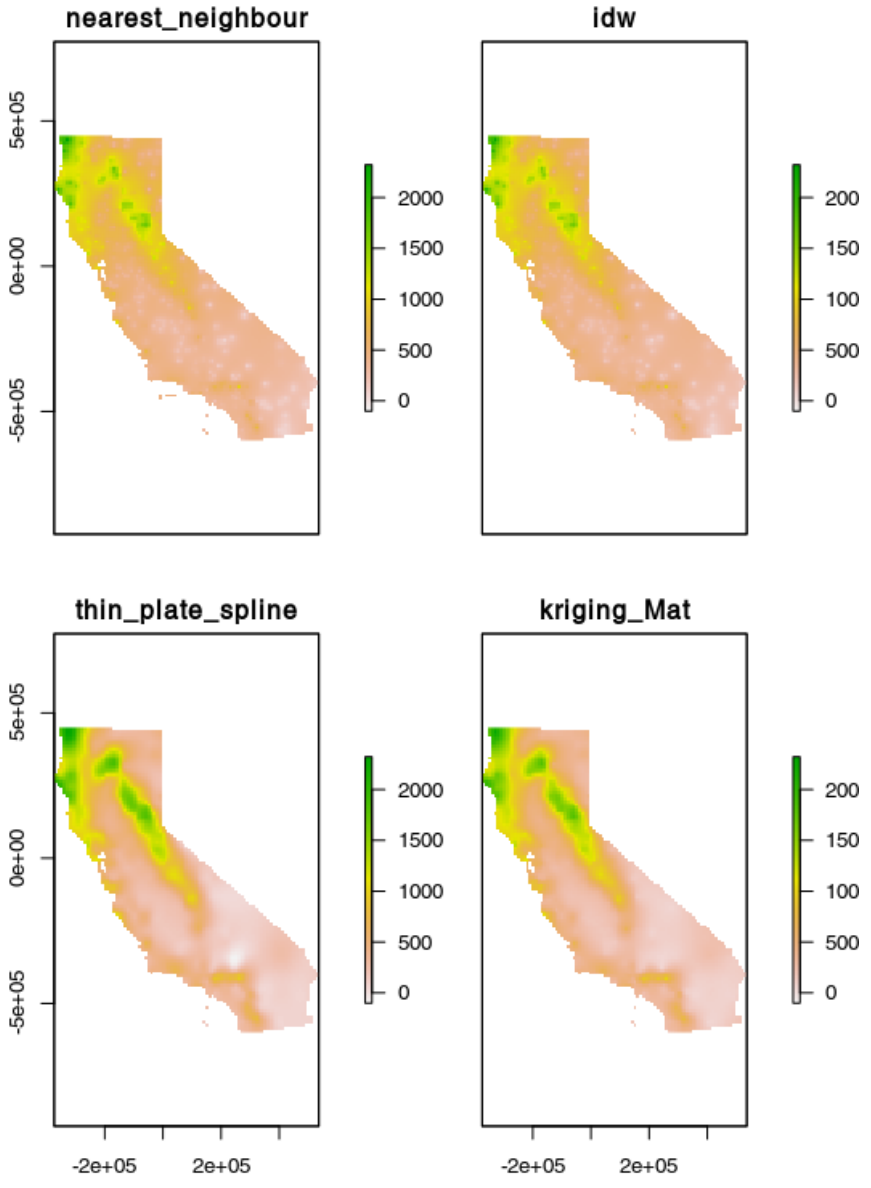
For an overview: online book "geocomputation with R"

R GIS - Spatial interpolation

Raster interpolation - check book R_Applied Spatial Data Analysis with R.pdf (chap. 8, 8.11)

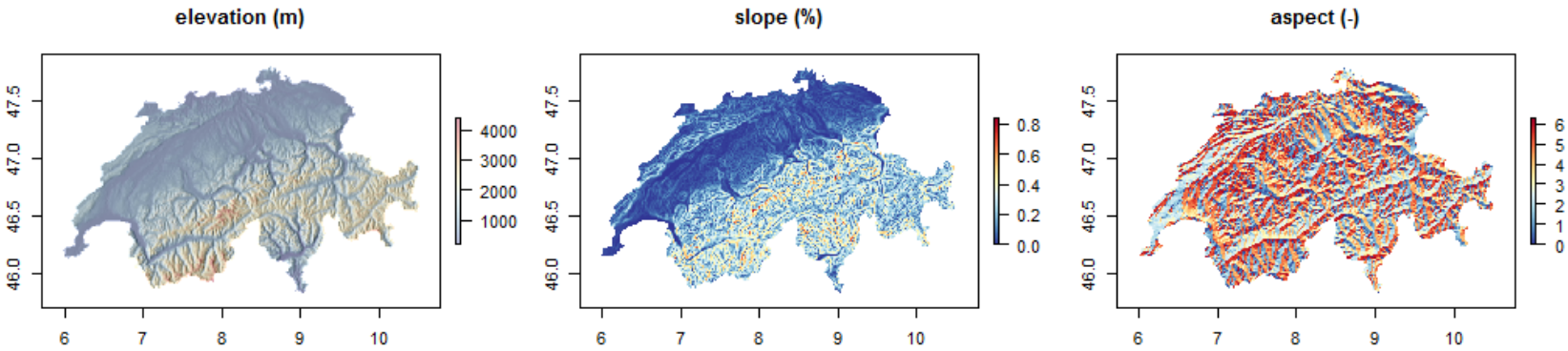
Adapted from here and here





R GIS - DEM processing

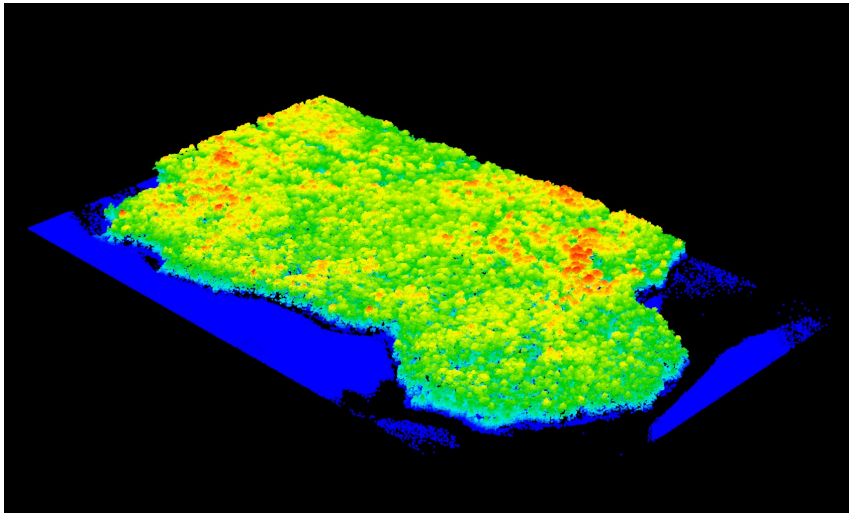
make a 4-4 pictures !!



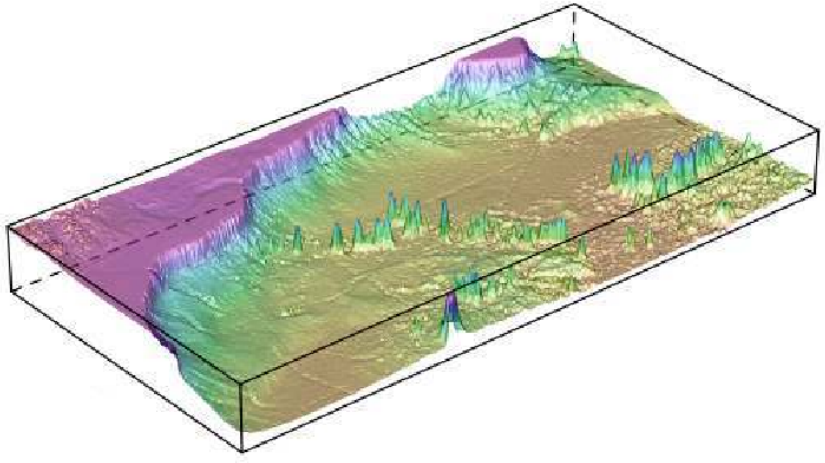
R GIS - 3D raster visualisation

R GIS - LiDAR & bathymetry

lidR: a R package for airborne LiDAR data manipulation and visualisation for forestry application (companion website)



`marmap`: a R package for making and analysing bathymetric (article)



R GIS - Feature projection

check graticules sf class

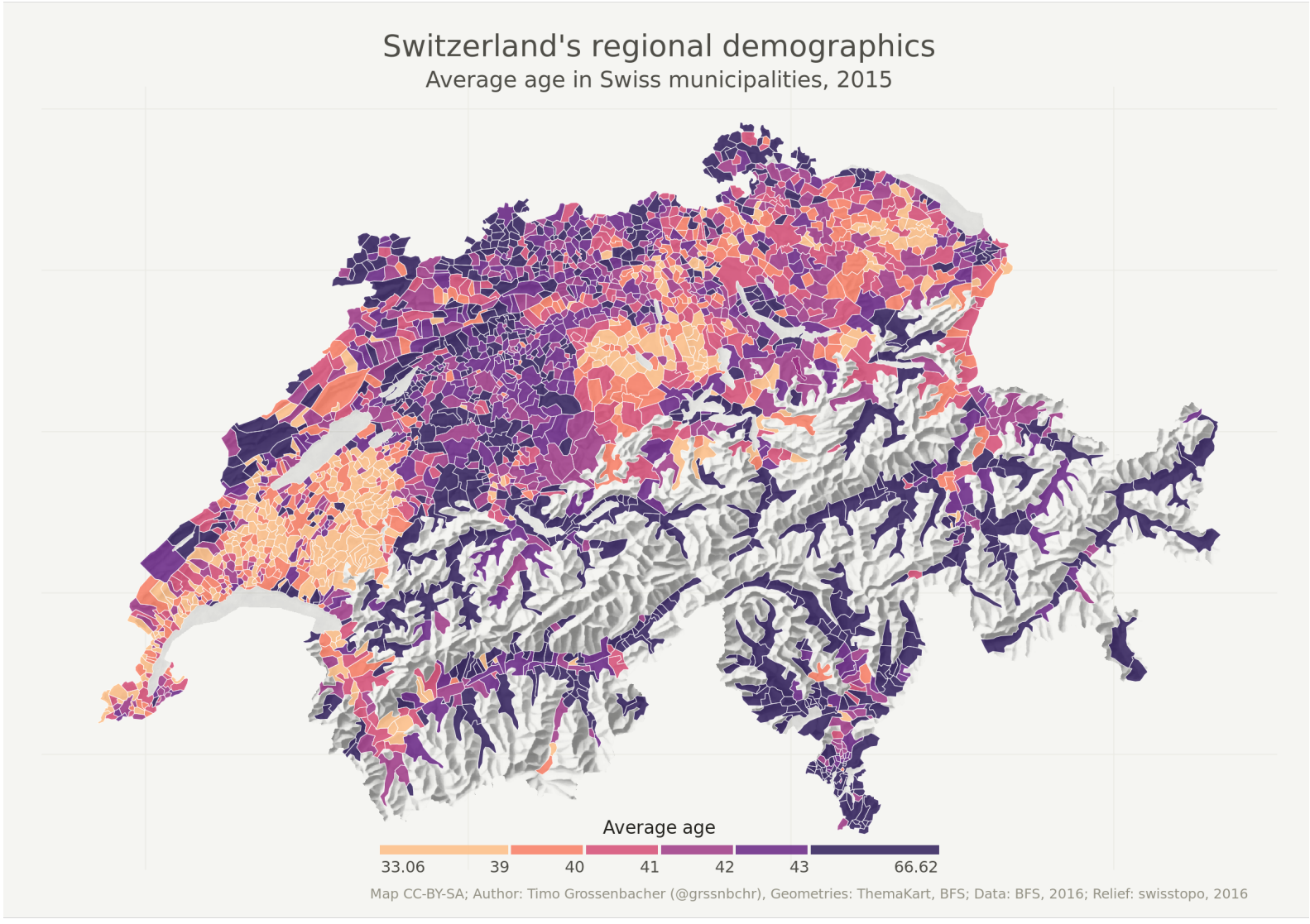
Mercator projection

Lambert Azimuthal Equal Area Projection

Lambert Conformal Conic Projection

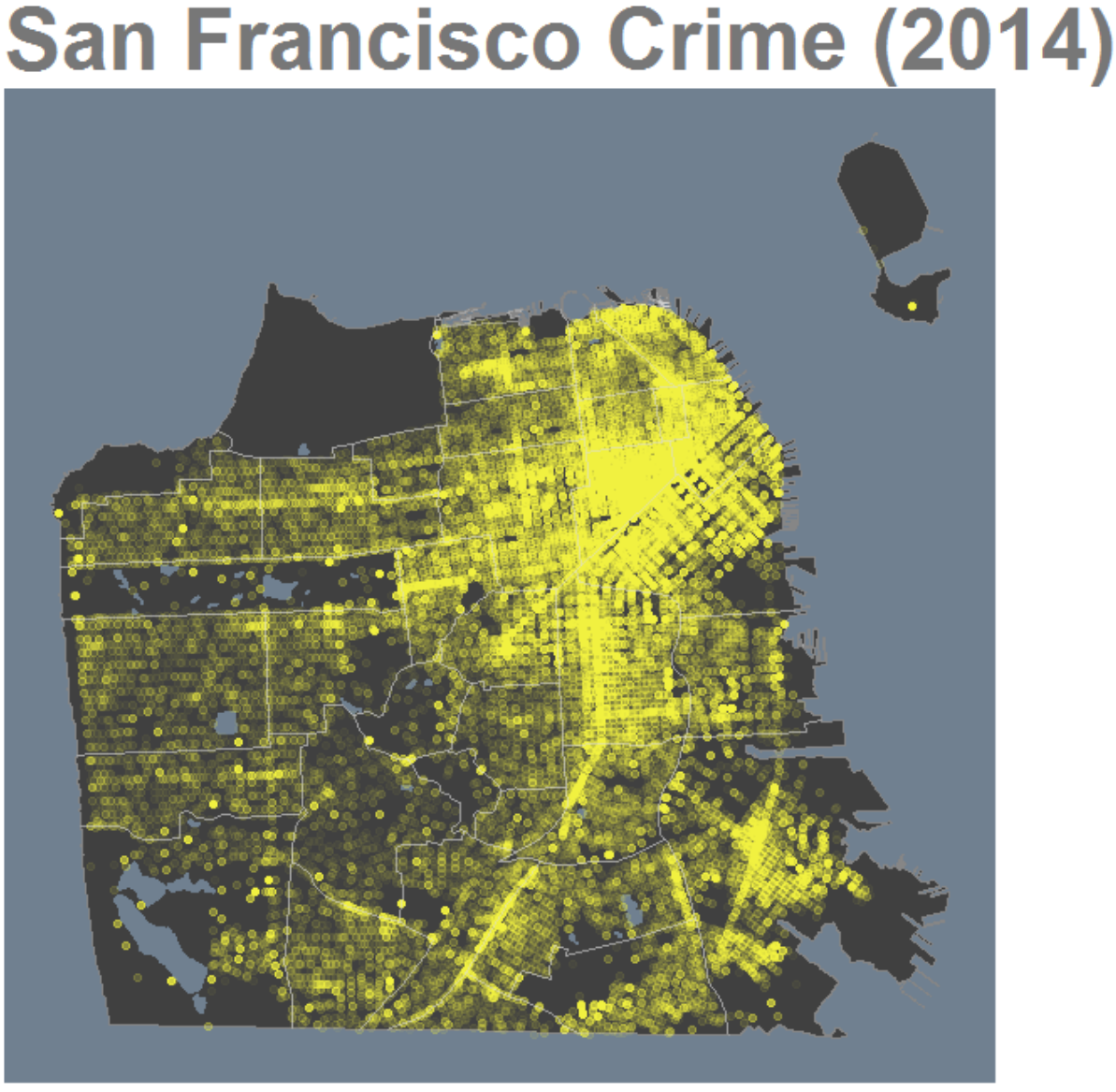
Albers Equal Area Conic

R GIS - Maps (1)



link

R GIS - Map (2)



link

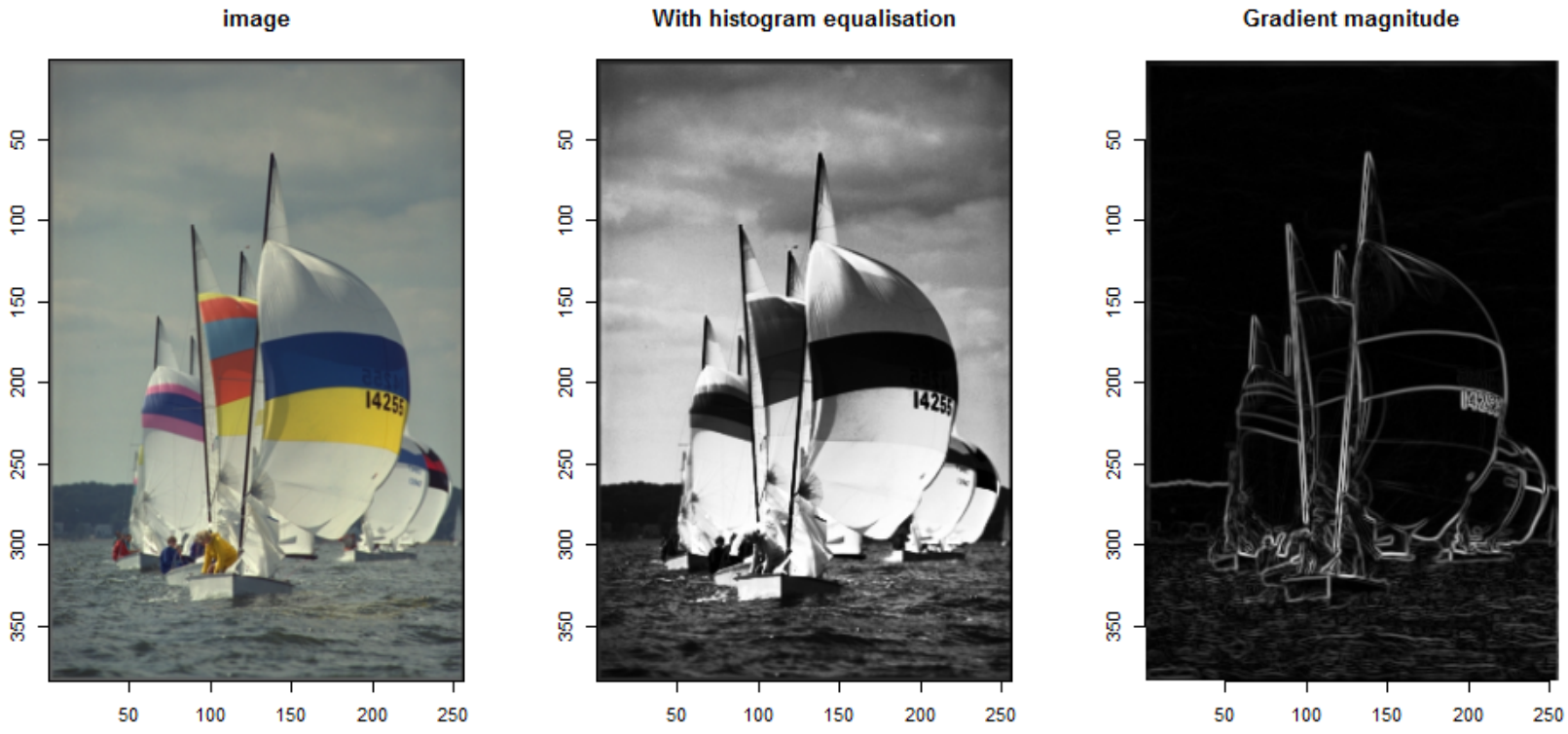
See also:

Image analysis and processing - resources

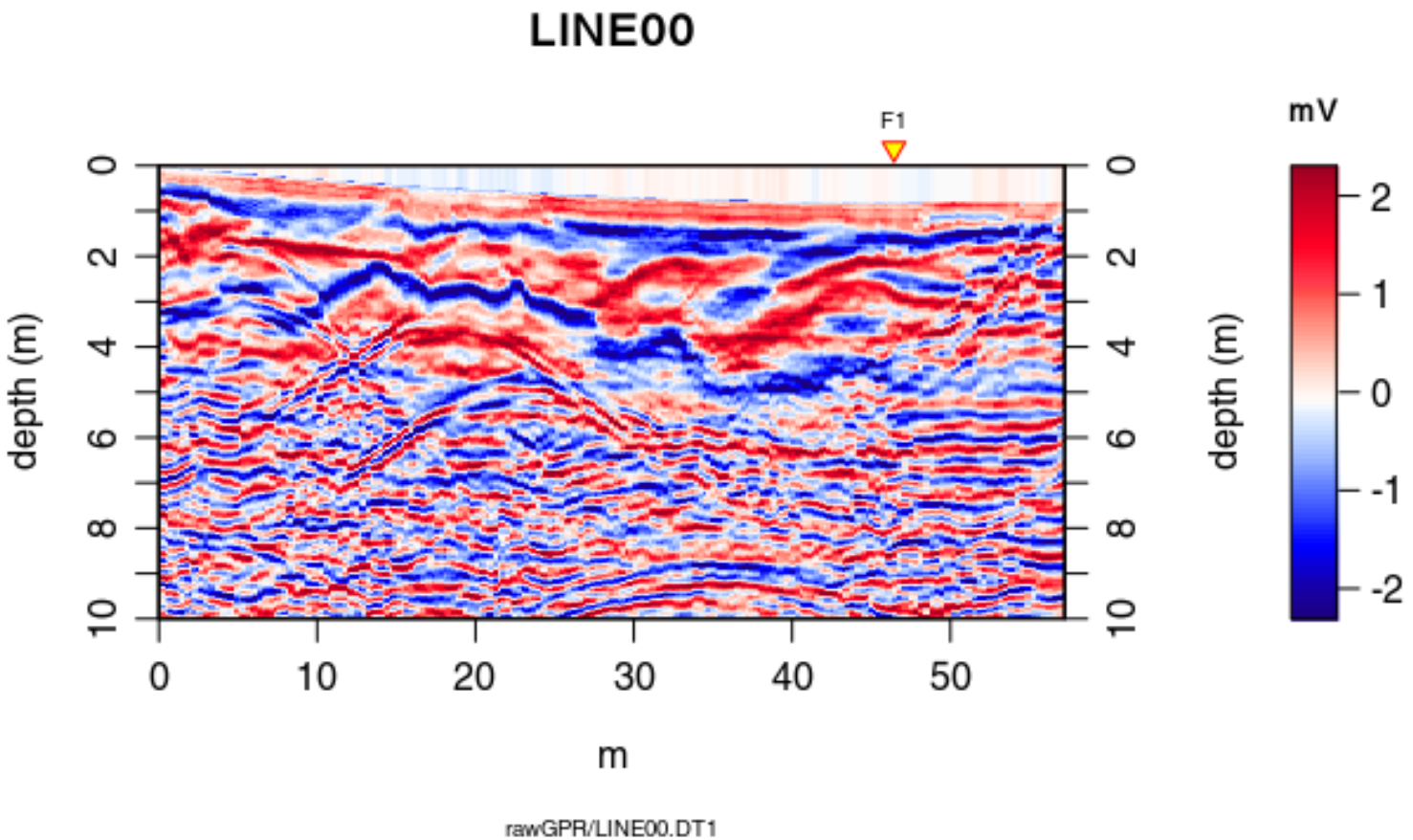
- **Package**
  - `imager`
  - `EBImage`
- **Tutorials/book**
  - check `imager` vignettes
  - `imager` project website
  - check `EBImage` vignettes
  - [Online presentation of `EBImage`] vignettes](https://www.bioconductor.org/help/course-materials/2015/BioC2015/BioC2015Oles.html)

Image analysis and processing

Example: `imager` package



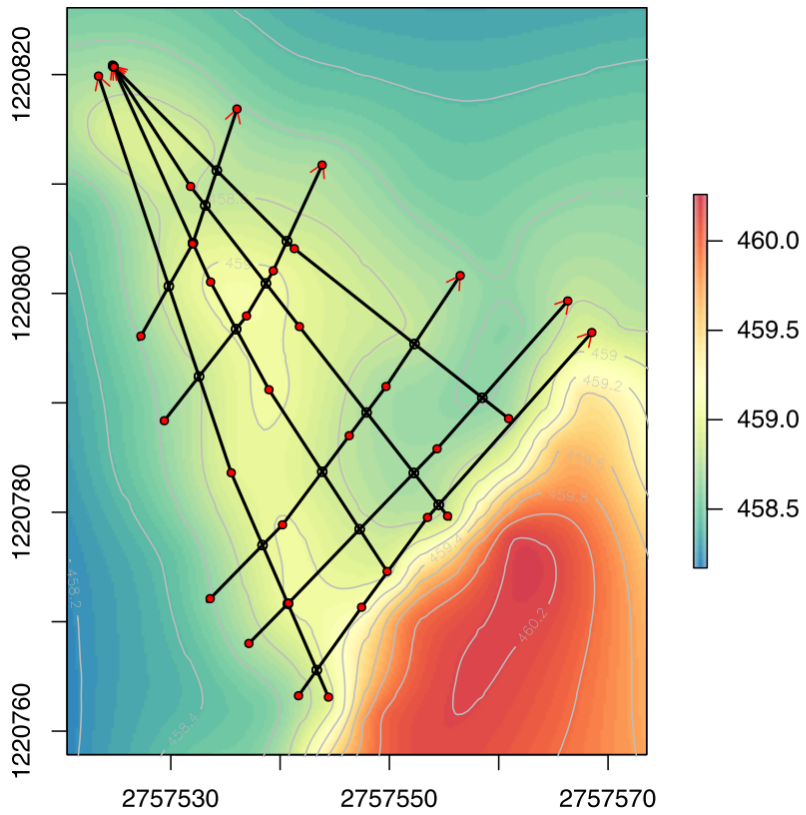
GPR data processing RGPR (1)

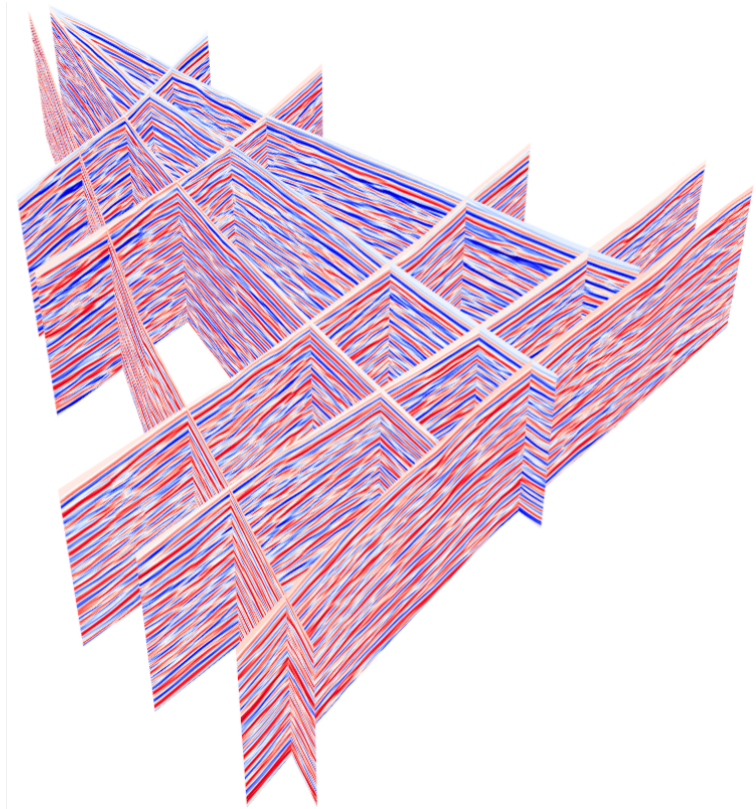


RGPR package

online tutorial

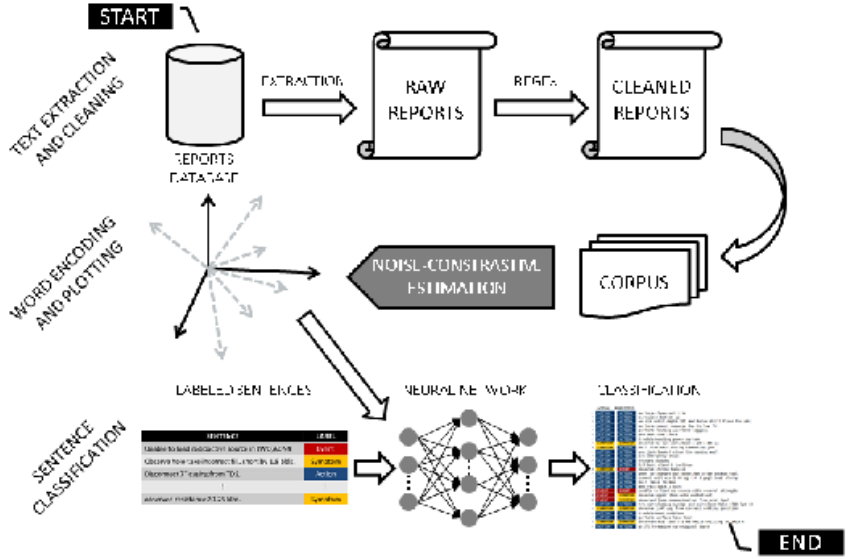
GPR data processing RGPR (2)





Database

- quality control: detect “outliers”, bad quality data, errors
- statistics, plots, reporting
- text analytic meet geosciences

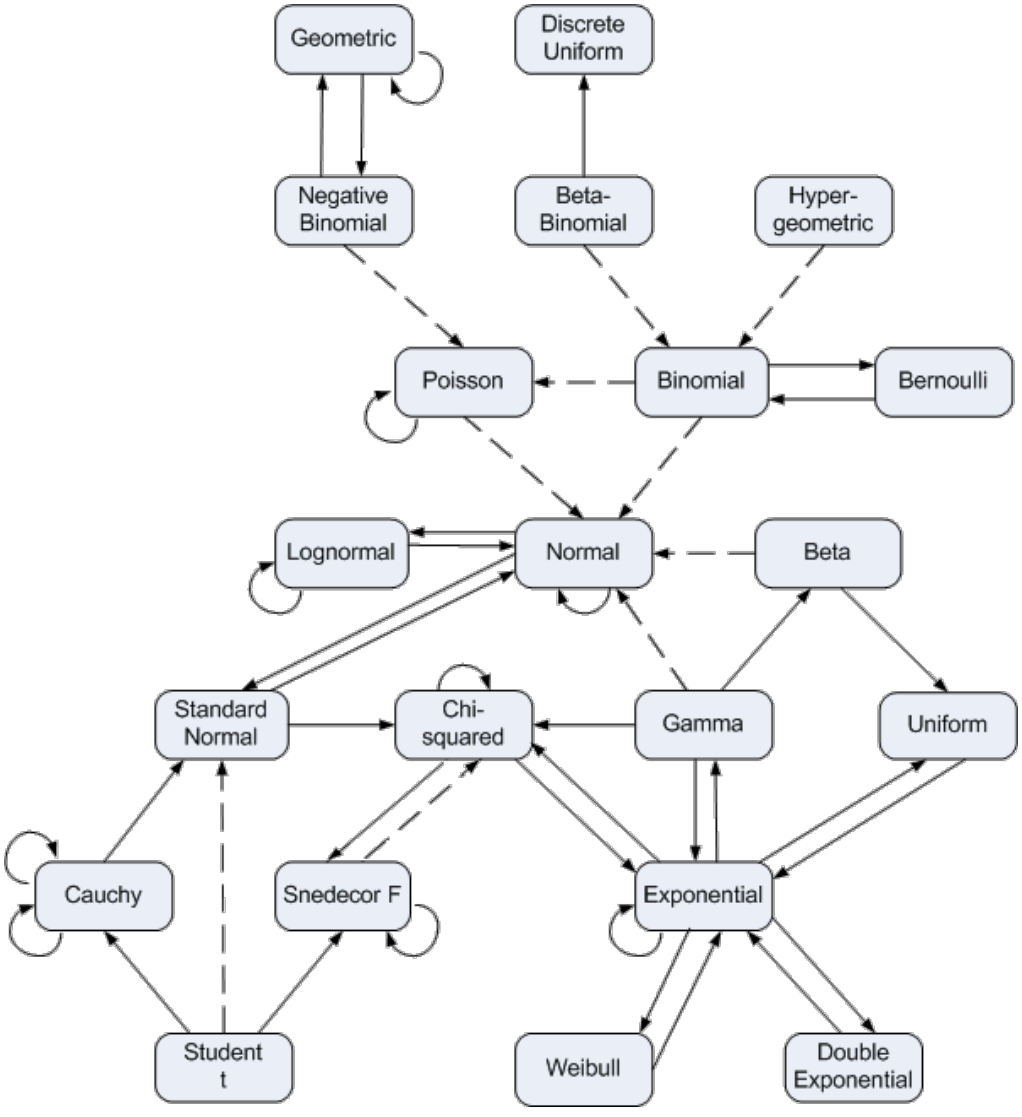


from arxiv.org/abs/1712.01476

- Regli et al. (2002) Interpretation of drill core and georadar data of coarse gravel deposits doi:10.1016/S0022-1694(01)00531-5

Modeling/simulations

Statistical/stochastic models - Probability models

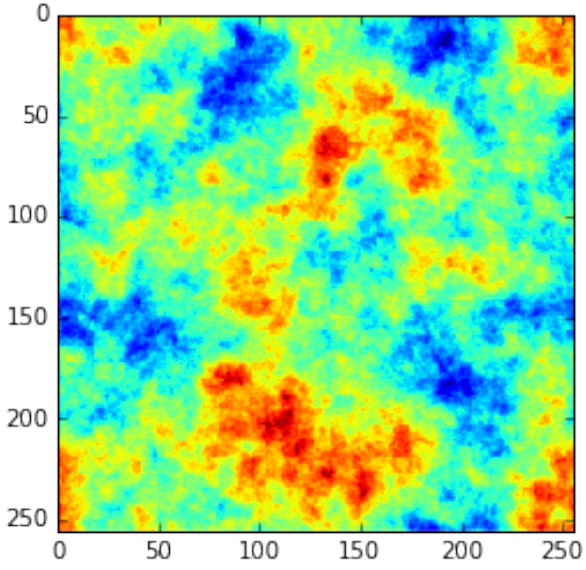


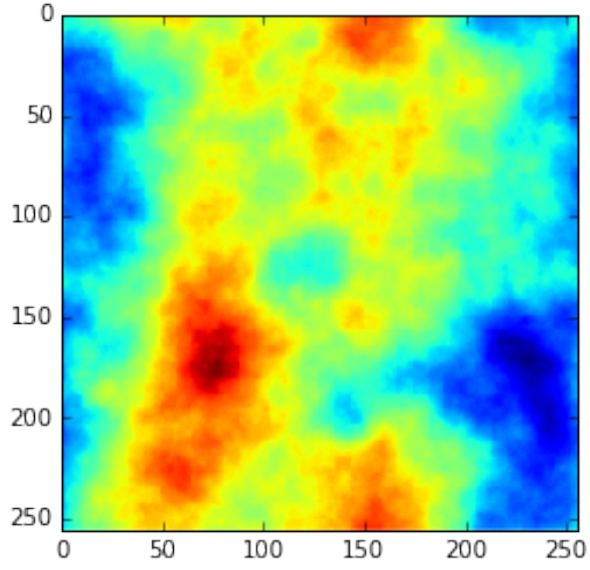
Bayesian modelling, MCMC, etc.

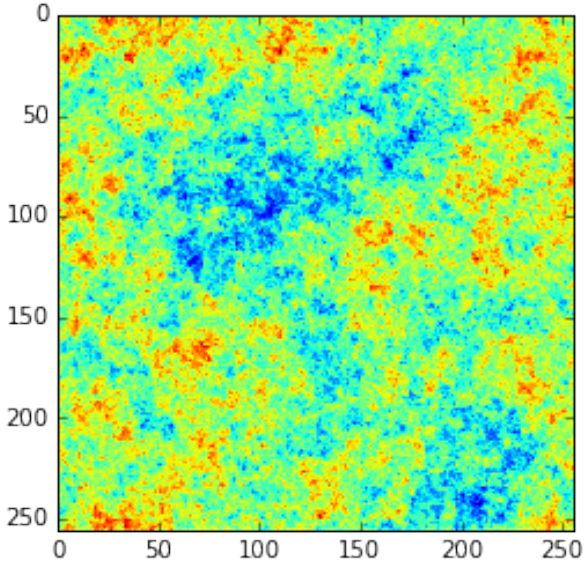
see CRAN Task View: Probability Distributions

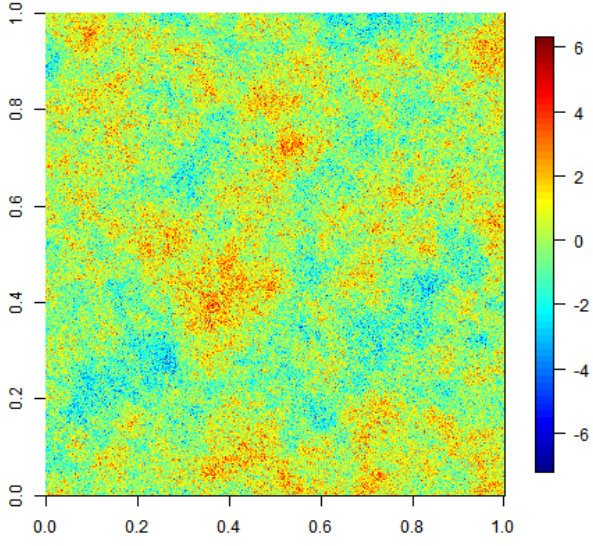
Statistical/stochastic models - Gaussian random fields

package `RandomFields`



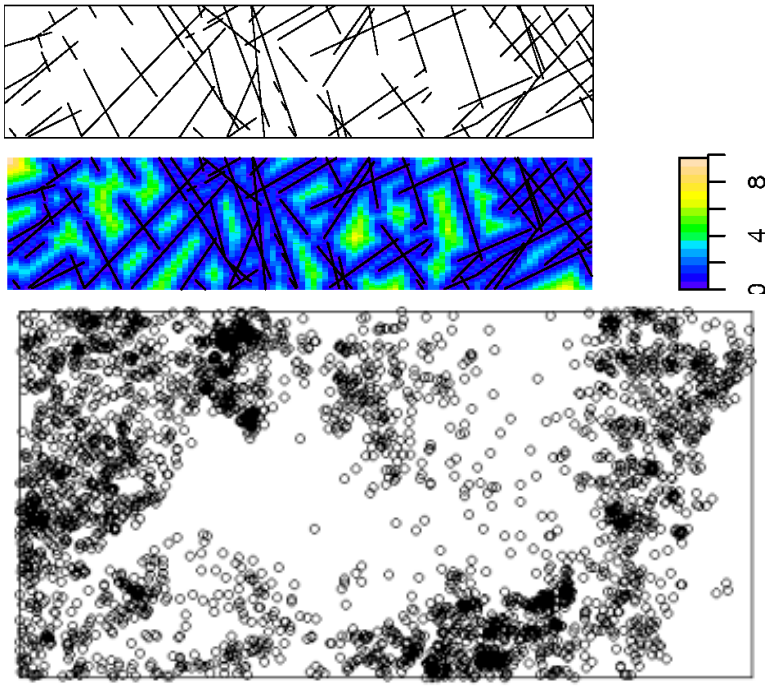


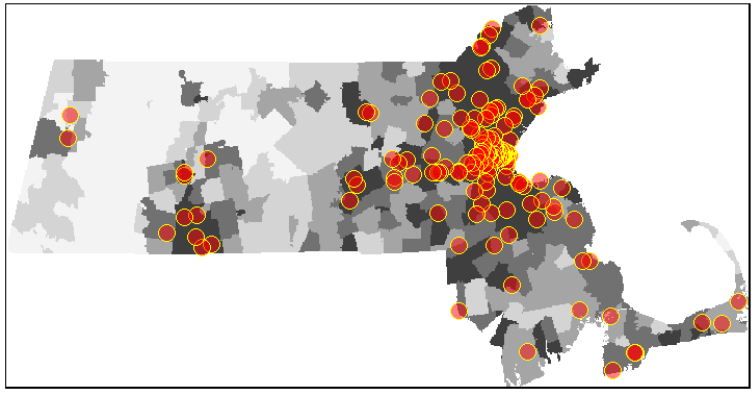


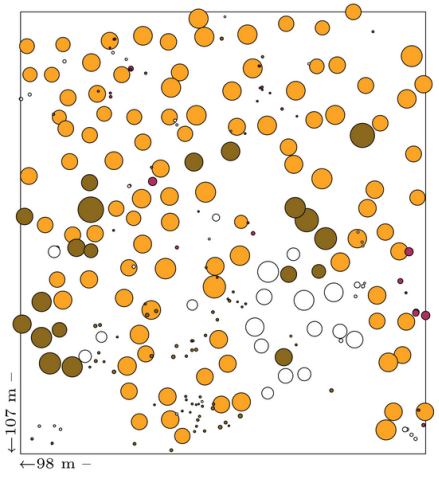


Statistical/stochastic models - Point processes

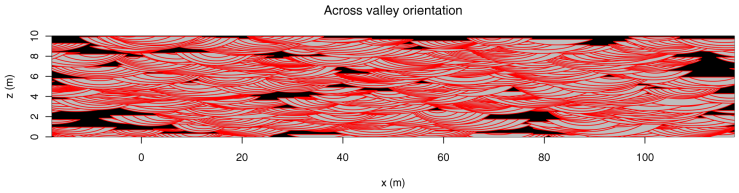
see `spatstat`

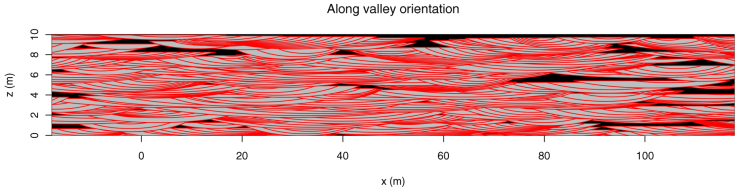


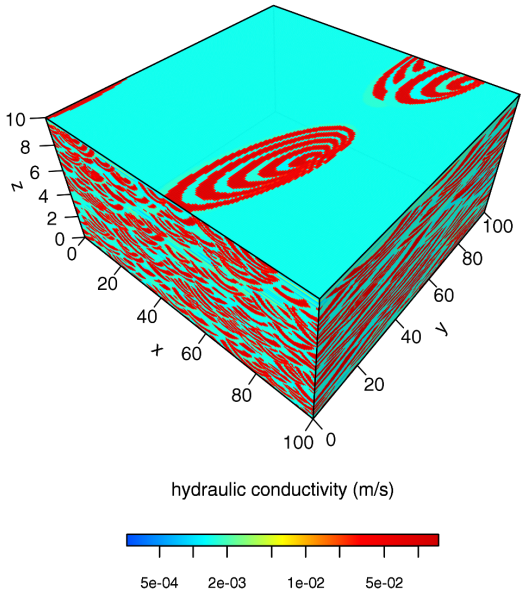


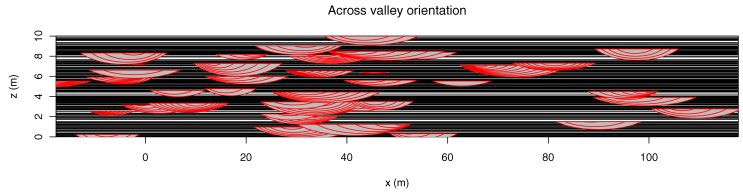


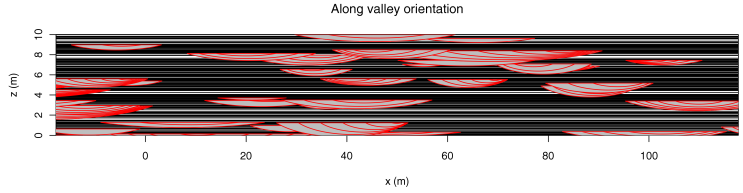
Coarse, braided river deposit model

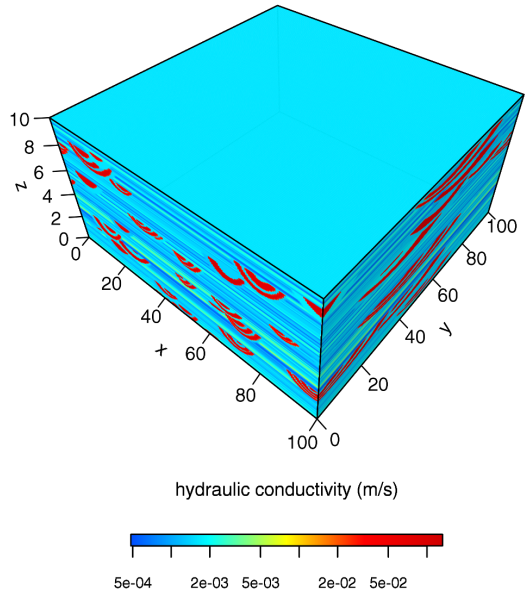












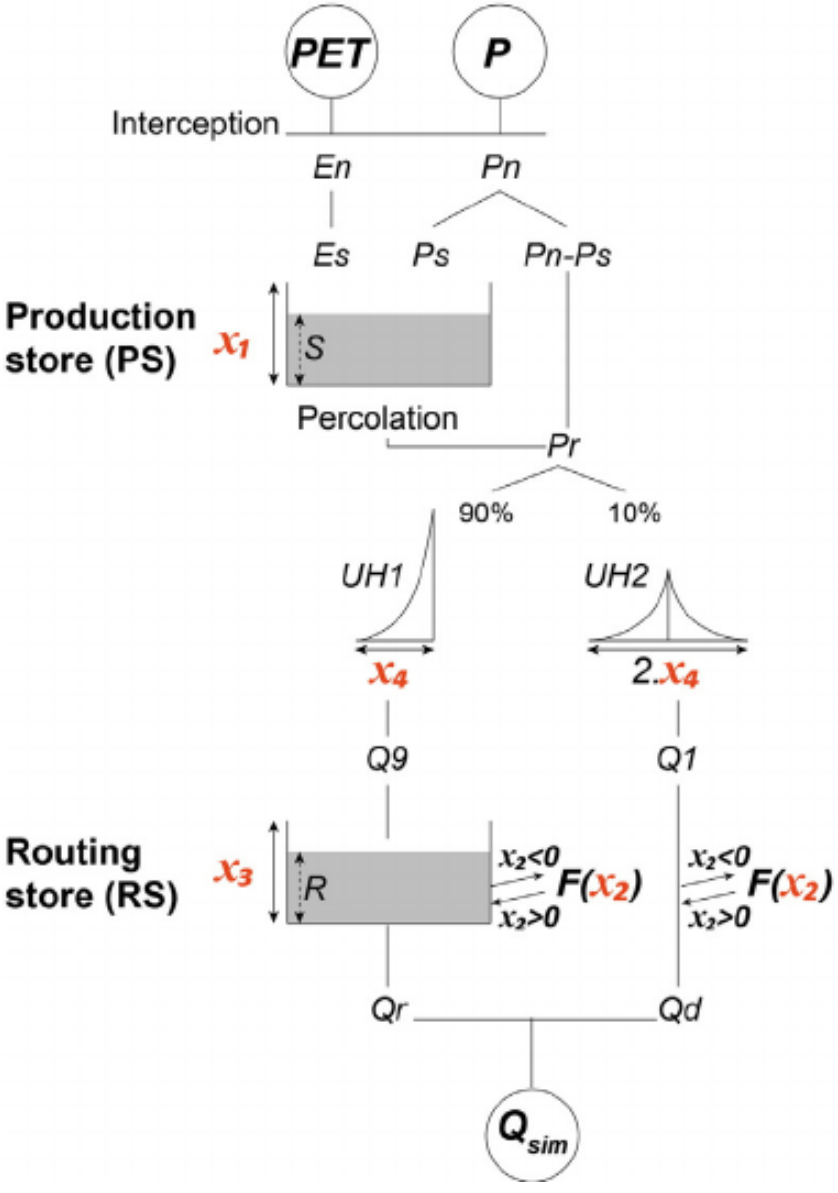
code and tutorial online

Catchment flow simulation

- `RRAWFLOW`: Rainfall-Response Aquifer and Watershed Flow Model rrawflow website
- `topmodel` and `dynatopmodel` packages
- package `ambhasGW` simulates groundwater fluctuations in response to precipitation and pumping: Sekhar et al. (2017)

Catchment flow simulation (1)

Package `airGR`, see tutorial on companion website



Simulation variables

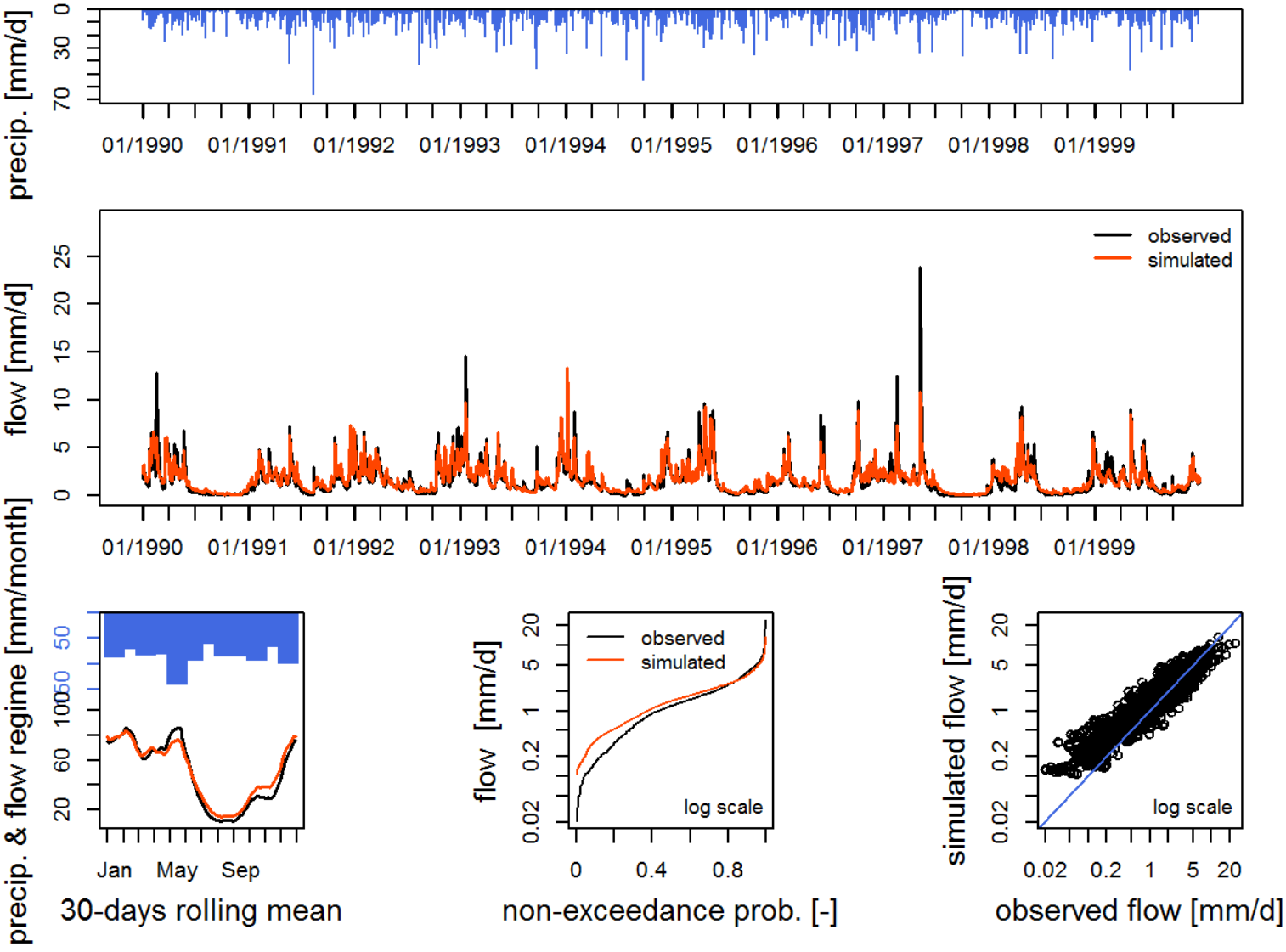
P: Precipitation
PET: Potential evapotranspiration
En: Net evapotranspiration capacity
Pn: Net precipitation
Es: Evaporation from the production store
Ps: Water filling the production store
S: Level in the production store
Pr: Water filling the routing store
UH1: Unit hydrograph 1
UH2: Unit hydrograph 2
Q9: Output of UH1
R: Level in the routing store
Q1: Output of UH2
Qr: Outflow of the routing store
F: Groundwater exchange term
Qd: Flow component from Q1 and F
Q_{sim}: Total simulated streamflow

Calibration parameters

x_1: maximum capacity of the PS (mm)
x_2: groundwater exchange coefficient (mm)
x_3: maximum capacity of the RS (mm)
x_4: time base of unit hydrograph UH1 (d)

Catchment flow simulation (2)

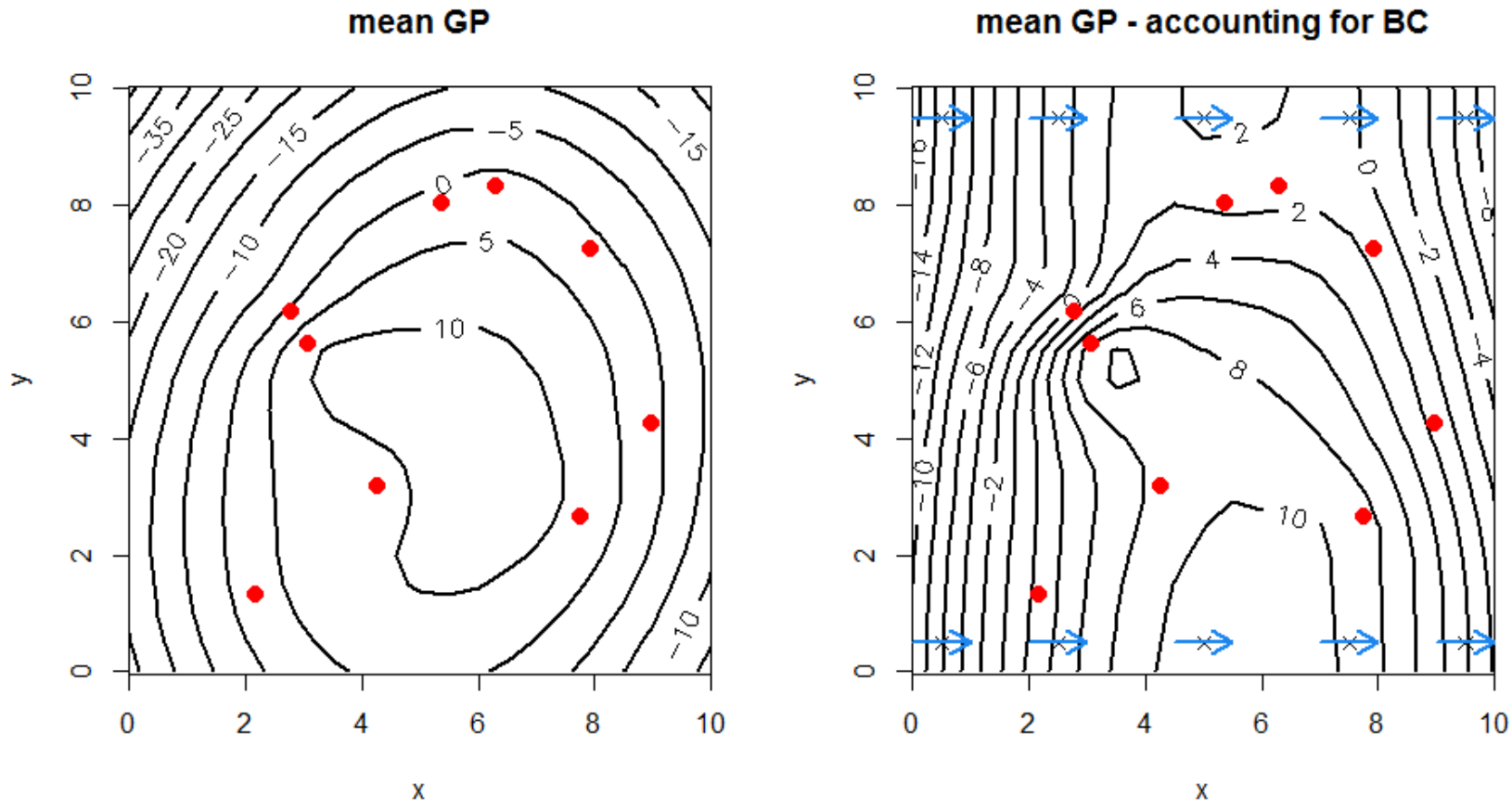
Package `airGR`, see companion website



physically based lumped model

Groundwater head interpolation

package `GauProMod`: Gaussian process (Kriging) with constraints on boundaries

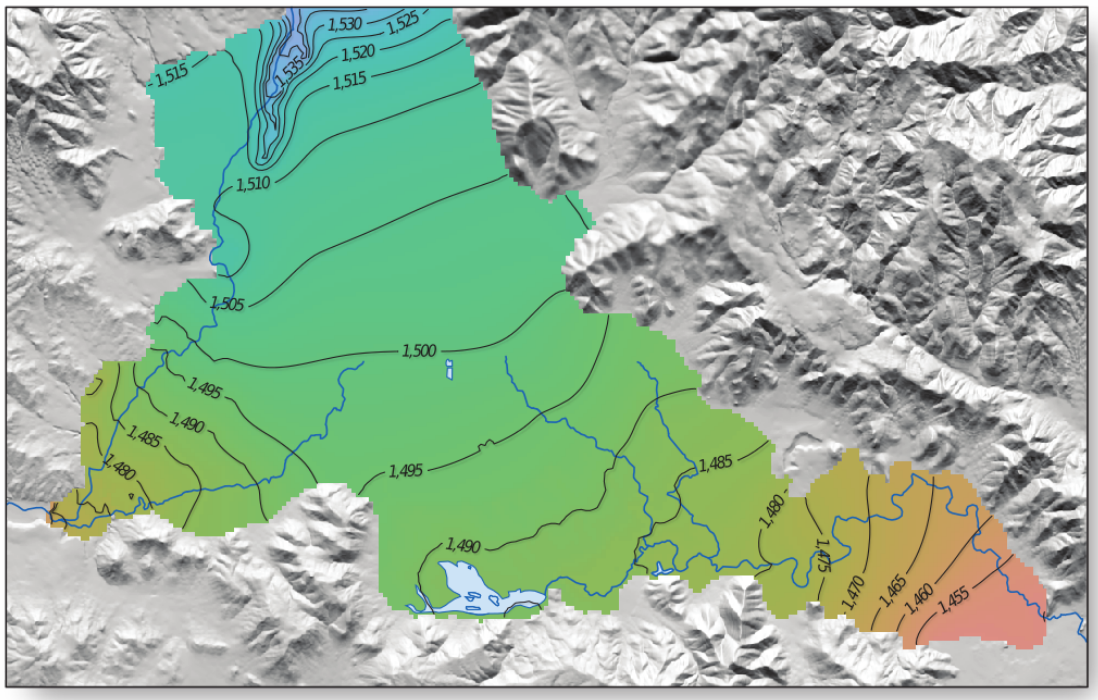


After Kuhlman and Igusquiza (2010)

Modflow - US-GS reproducible report

Groundwater-flow model for the Wood River Valley aquifer system, south-central Idaho

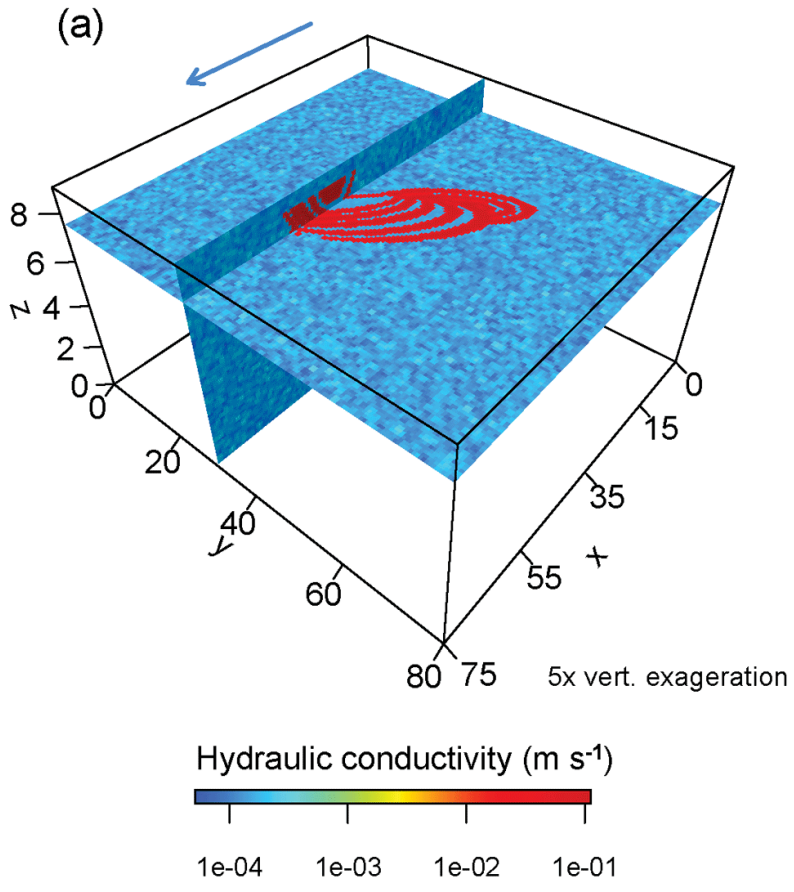
Use R for GIS operation to set up a groundwater model (MODFLOW).

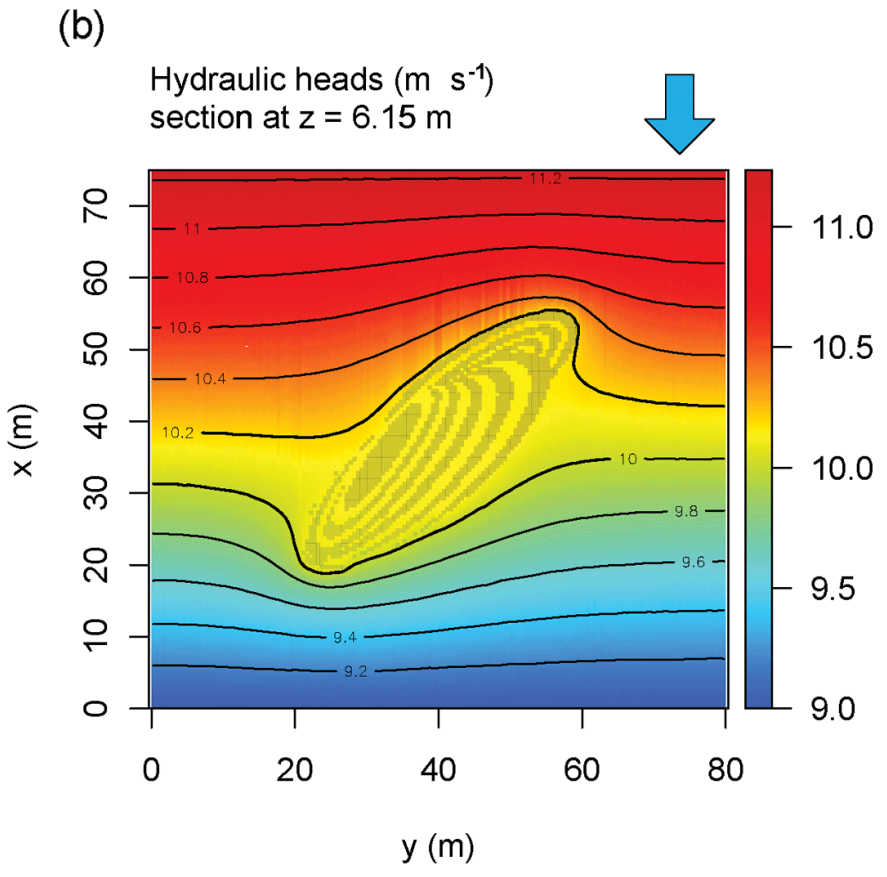


github.com/USGS-R/wrv

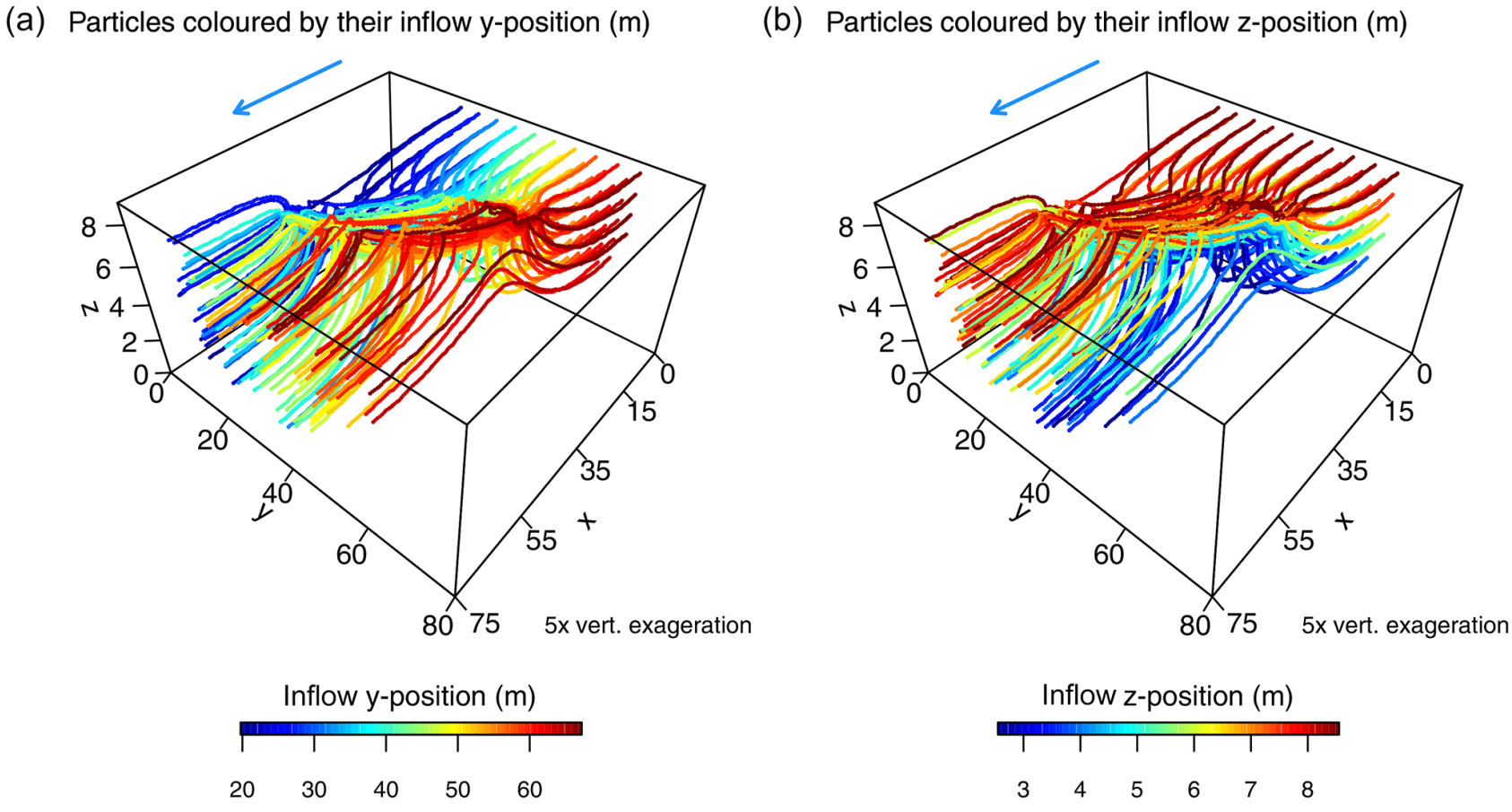
Modflow/modpath - subsurface flow mixing (1)

personal code based on github.com/USGS-R/wrv

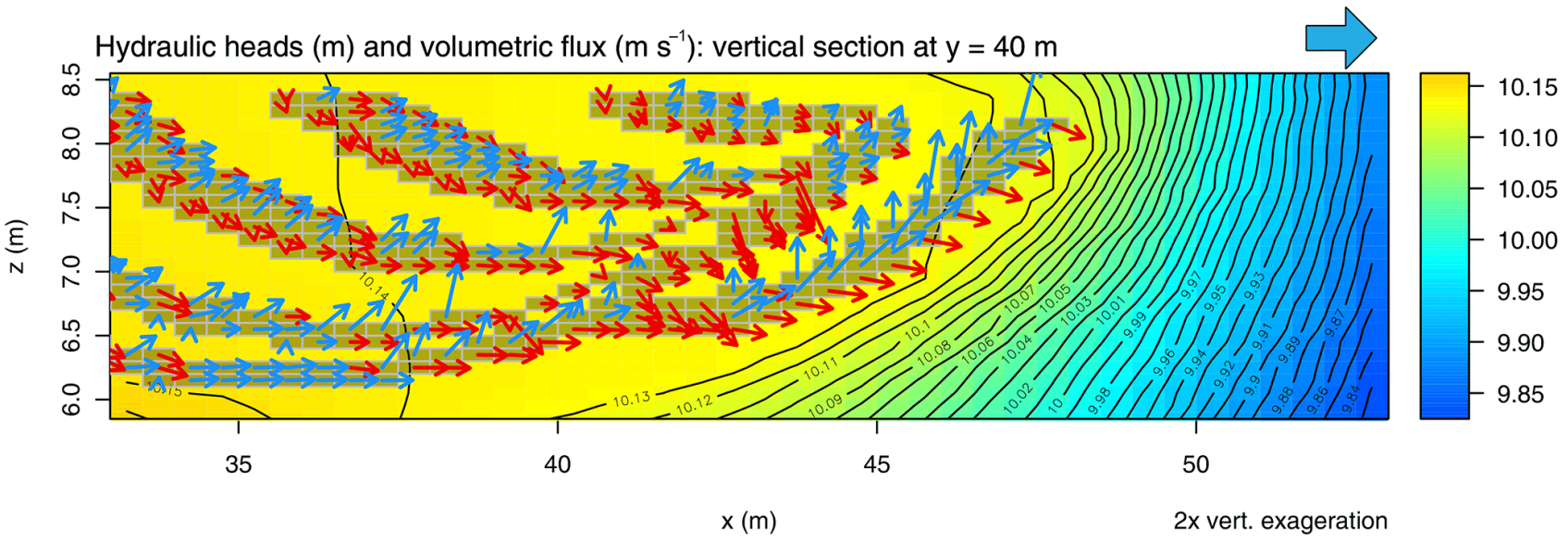




Modflow/modpath - subsurface flow mixing (2)



Modflow/modpath - subsurface flow mixing (3)

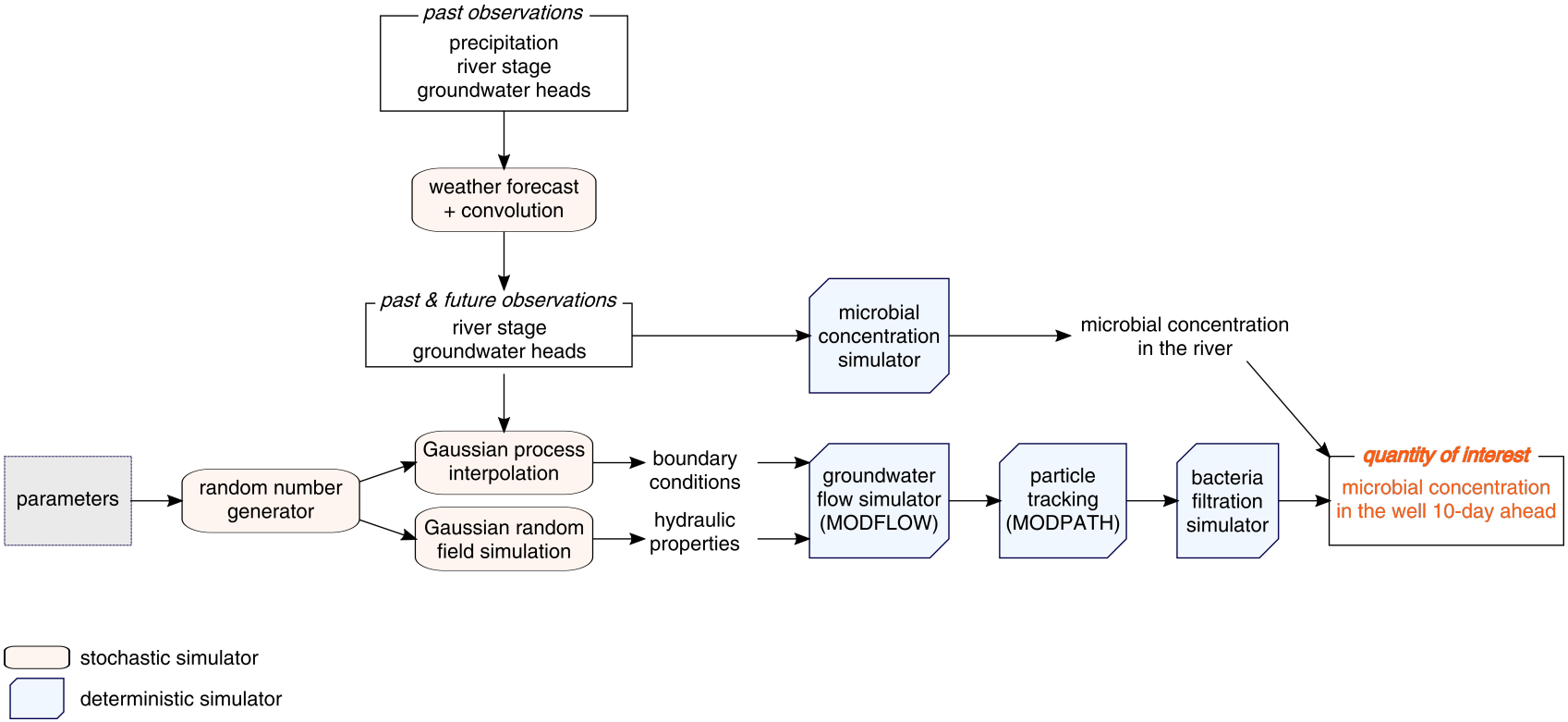


Modflow/modpath -stochastic simulation (1)

(`gwModBac` personal code based on github.com/USGS-R/wrv

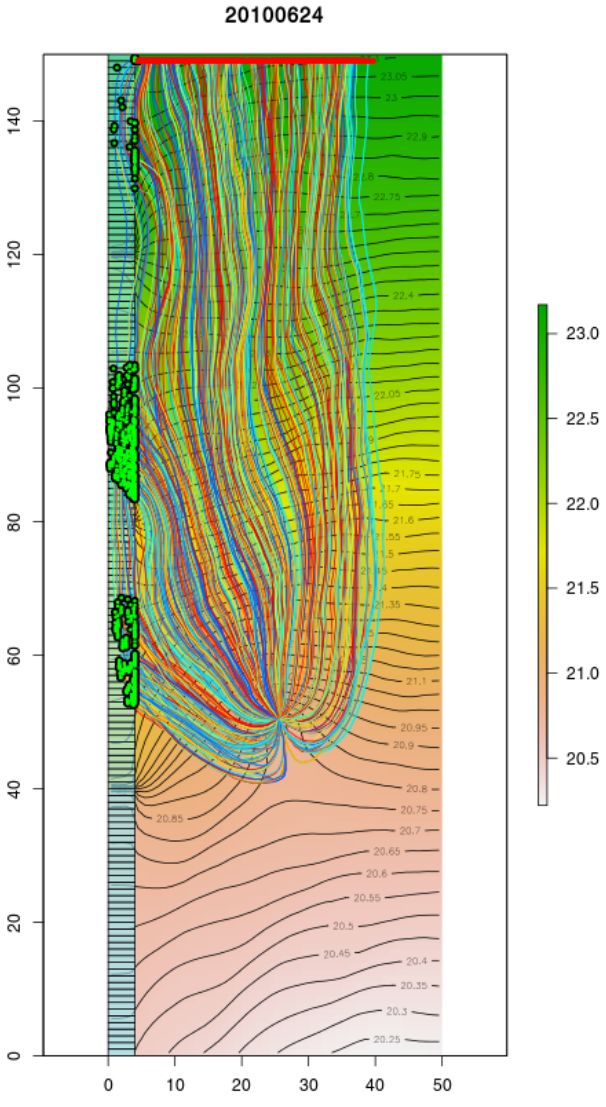
Groundwater flow simulation and particle tracking to forecast microbial concentration in a drinking water extraction well.

Workflow

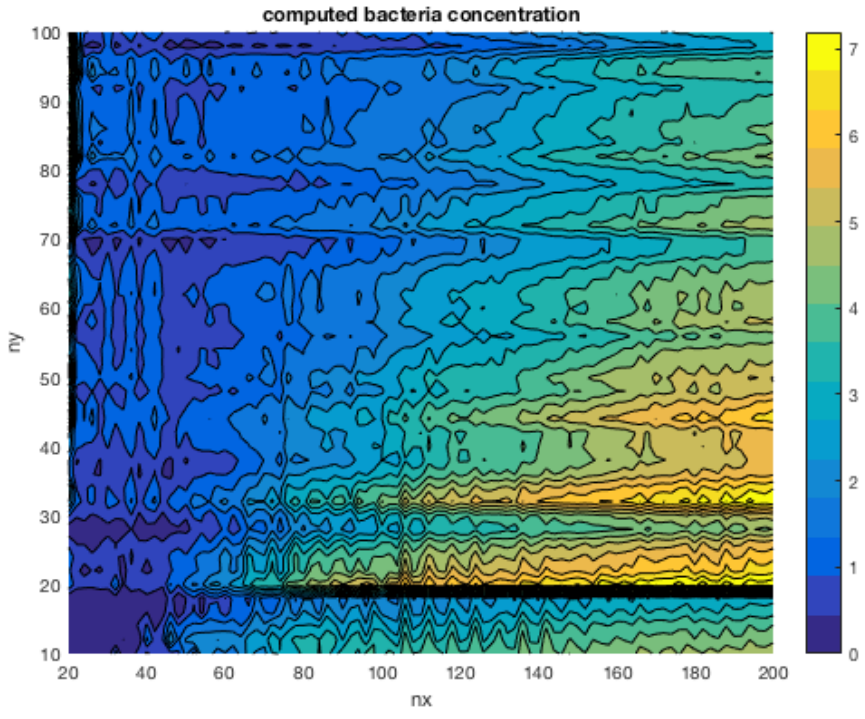


Modflow -stochastic simulation (2)

Bacteria pathways



Bacteria concentration as a function of grid size



Other applications:

- uncertainty quantification (MCMC)
- sensitivity analysis
- teaching

Reporting

Ressources - reporting

Two philosophies:

- `ggplot` and `ggplot2` and theirs cousins
- or base `graphics`, `plot3D`, etc.

3D plots:

- `rgl` for 3D interactive plots (based on OpenGL)
- `rasterVis` for 3D interactive plots of raster (based on OpenGL)
- `plot3D` for 3D static plots

publication-quality figures (see also package `ggplot2`)

See R-graph gallery

large graphics choice

- 2D and 3D plots
- static and interactive plots

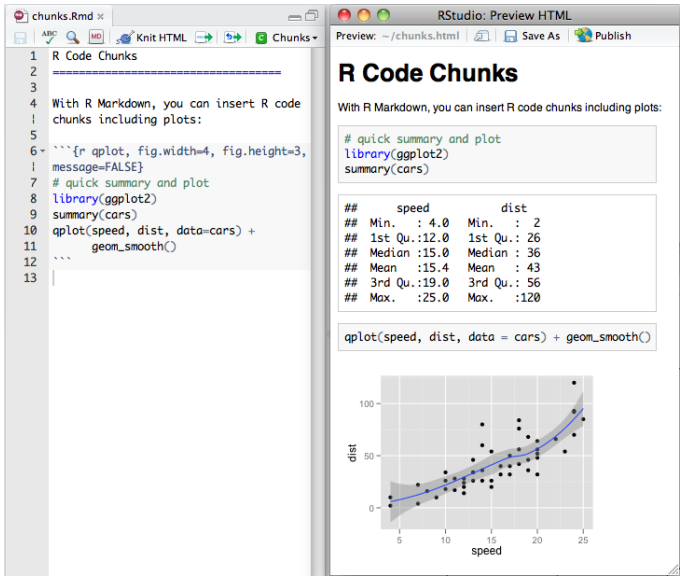
Export

- png, jpeg, gif
- pdf
- svg
- html (3D)
- latex (?)

Report, presentation, book

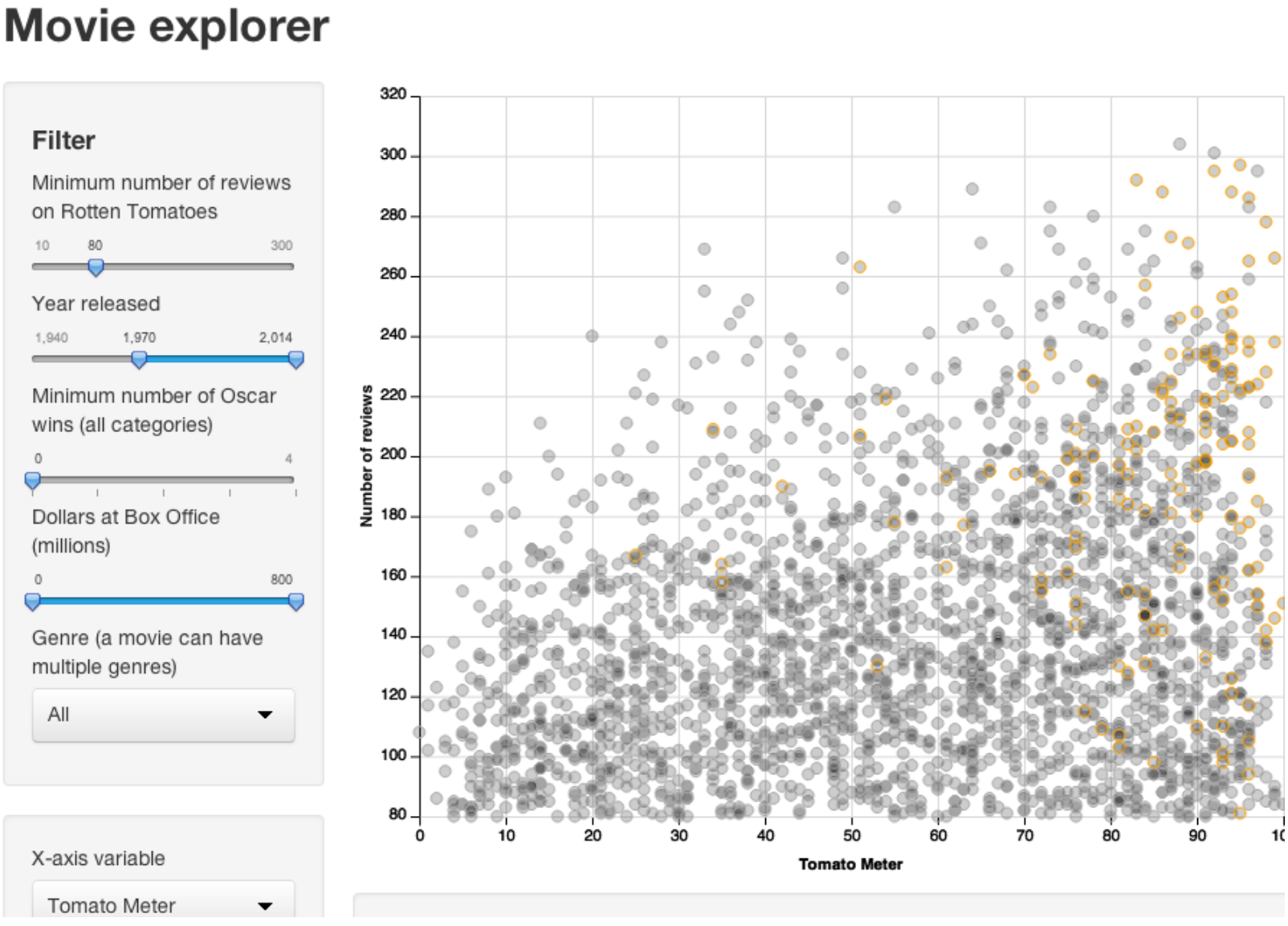
bookdown

Report/presentation/book: HTML/PDF (packages `RMarkdown` and `knitr`)



Web applications

interactive web apps (package `shiny`)



R package

- combine several codes in a package
- include data in the package

Tools

RStudio

- code, figures (export), help, etc.
- package creation and development
- report (R markdown)

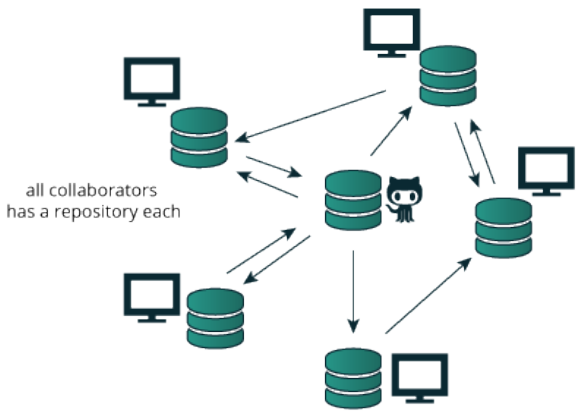
Git & Github

git

a free and open source distributed version control system

github

a web-based hosting service for version control using git



Github public/private repository:

- host code with version contral
- several branches (develop)
- wiki
- project board (task management)
- host project website

Reproducible research

from Karl Broman

0. Separate the raw data from everything (and don't modify them)
1. Organize your data & code
2. Everything with a script
3. Automate the process (GNU Make)
4. Turn scripts into reproducible reports
5. Turn repeated code into functions
6. Create a package/module
7. Use version control (git/GitHub), no more "really_true_final_2EH5b.doc"
8. Pick a license, any license

What's next?