

Programming language R

Capabilities for data transformation, analysis,
visualization and dynamic document building

Didzis Elferts



Interreg
Estonia-Latvia
European Regional Development Fund



EUROPEAN UNION

WaterAct

Joint actions for more efficient management
of common groundwater resources

Outline

- R software and language
- RStudio
- R capabilities

About me

- Dr.biol., assoc.prof. Didzis Elferts
- Faculty of Biology, University of Latvia
- More than 10 years experience in R
- University courses: biometry (statistics) for biology bachelor and master students, geography/environmental science doctoral students
- e-mail: didzis.elferts@lu.lv

What is R?

- R is a programming language and open-source software
- R was started as alternative implementation of S language
- Version 1.0 was released on 29th February 2000
- Version 4.0.3. (Bunny-Wunnies Freak Out) was release on 10th October 2020

Source: [https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))

Pros and cons of R

Pros:

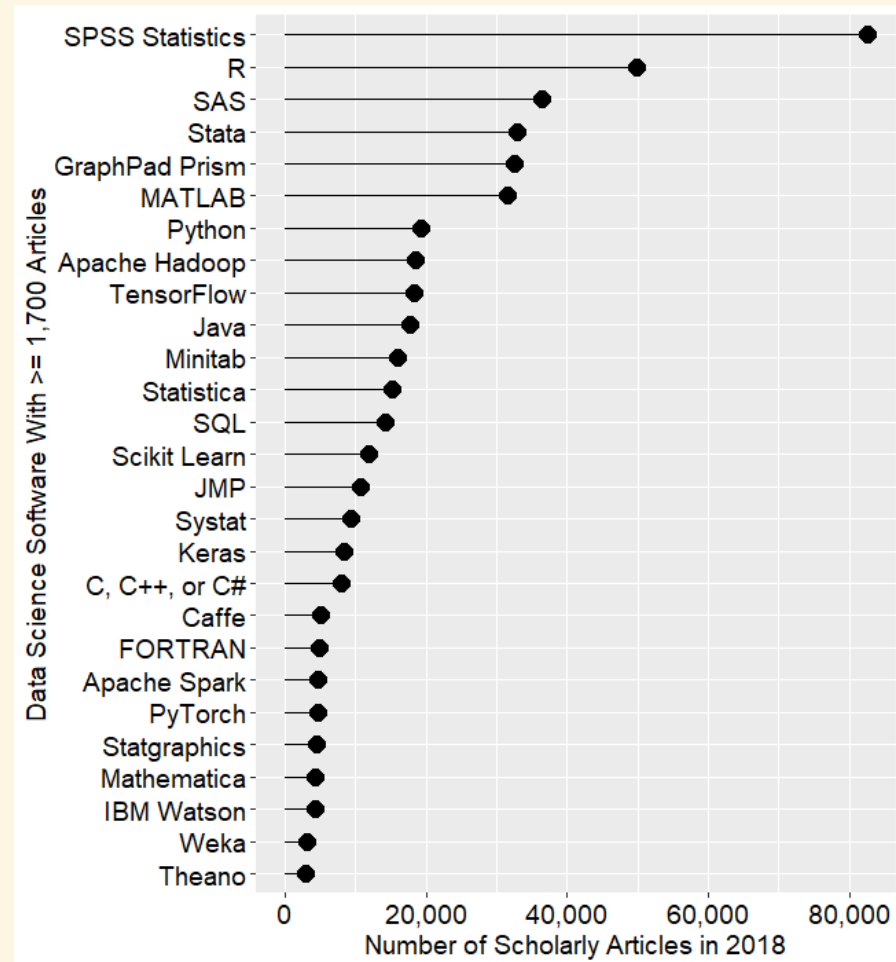
- Open-source, regularly updated, still developing program;
- Works on main platforms: Windows, MacOS, Linux
- Different statistical methods implemented, control over parameters
- Excellent graphical capabilities
- Own functions and R packages
- Development of interactive visualizations, web-applications
- Ideal to implement reproducible research

Pros and cons of R

Cons:

- Slow "learning" pace
- Partly - comandline program
- Sometimes hard to find necessary information/package/function

R popularity



Source: <http://r4stats.com/2019/04/01/scholarly-datasci-popularity-2019/>

Usefull links

- R program homepage <http://www.r-project.org/>
- RStudio homepage <http://www.rstudio.com>
- YouTube channel with tutorials <https://www.youtube.com/playlist?list=PLcgz5kNZFCkzSyBG3H-rUaPHoBXgijHfC>
- Q&A page Stack overflow <http://stackoverflow.com/>
- Search in R packages <http://www.rdocumentation.org/>

R packages

- Base R - only small part of statistical analysis, base graphics
- Additional capabilities through R packages (libraries)
- Developed by users, hosted on CRAN (official), github or internally
- CRAN packages: 16517 (11.11.2020.)

R packages

- `kwb.hantush` - Calculation of Groundwater Mounding Beneath an Infiltration Basin
- `GWSDAT` - GroundWater Spatiotemporal Data Analysis Tool
- `hydrogeo` - Groundwater Data Presentation and Interpretation
- `streamDepletr` - Estimate Streamflow Depletion Due to Groundwater Pumping
- `dataRetrieval` - Retrieval Functions for USGS and EPA Hydrologic and Water Quality Data

RStudio

- RStudio is a company developing free and open tools for R, as well as, enterprise-ready professional products
- Software: RStudio IDE, RStudio Server, Shiny Server
- Cloud: RStudio Cloud, shinyapps.io
- R packages: tidyverse, ggplot2, dplyr, tidyr, purrr, stringr, shiny, rmarkdown, flexdashboard, sparklyr, tidymodels, reticulate, plumber,

Source: <http://www.rstudio.com>

Thinks to remember

- To use more than one processor core, additional packages and functions needed. Can be extended to use computer clusters, cloud computing, GPU computing
- R stores all data used for calculations in RAM

Data sources

- Most data formats supported (mainly additional R packages needed) - txt, cvs, xlsx, sav, json, NetCDF, ...
- Direct download from the webpages
- Connection to databases (also to password protected), data filtering - can use SQL commands or R language
- Data from loggers - if there is no function and data have the same pattern, we can make it!

Data sources

- Think about the possibility that data will be used in other programs
- Tidy data:
 - Each variable must have its own column
 - Each observation must have its own row
 - Each value must have its own cell

<https://r4ds.had.co.nz/tidy-data.html>

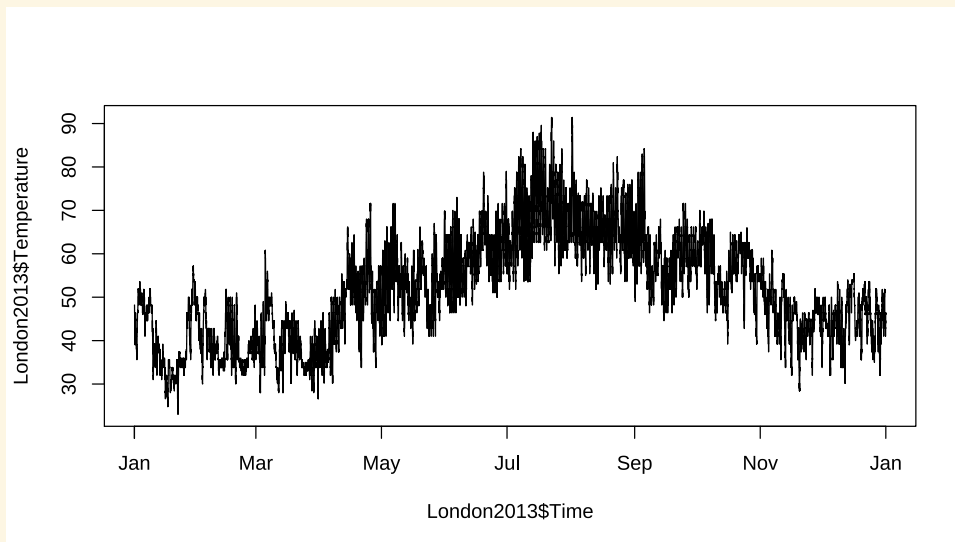
Data transformation

- One of directions - `tidyverse` packages
- Data cleaning, change of layout (wide format to long format)
- Data summarizing
- Subsetting, filtering, joining of tables, etc.

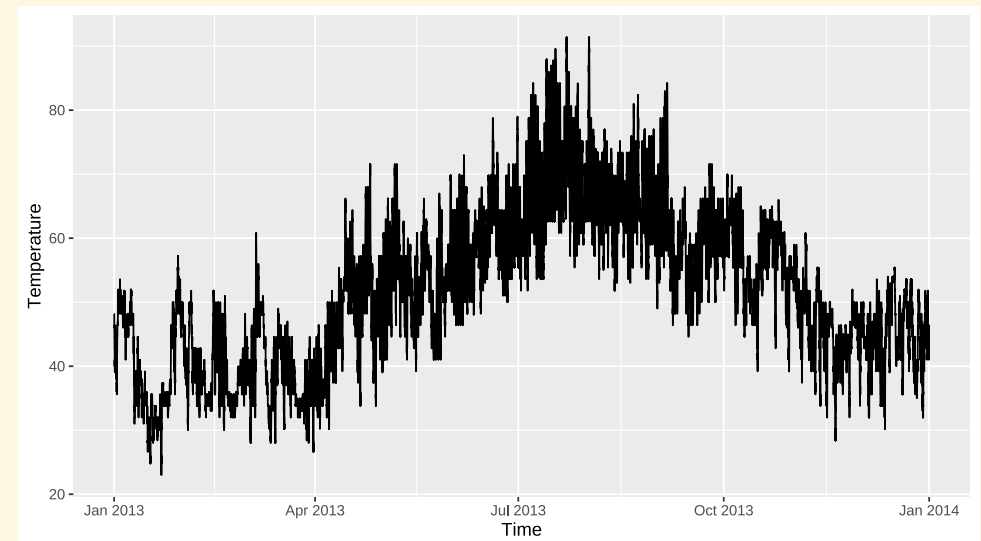
Data visualisation

- Different systems - base, ggplot2, lattice. New possibilities - interactive plots, leaflets, etc.

```
library(weatherData)
library(lubridate)
data("London2013")
London2013$Time <- ymd_hms(London2013$Time)
plot(London2013$Time, London2013$Temperature, type = "l")
```



```
library(ggplot2)
ggplot(London2013, aes(Time, Temperature)) +
  geom_line()
```



Data visualisation

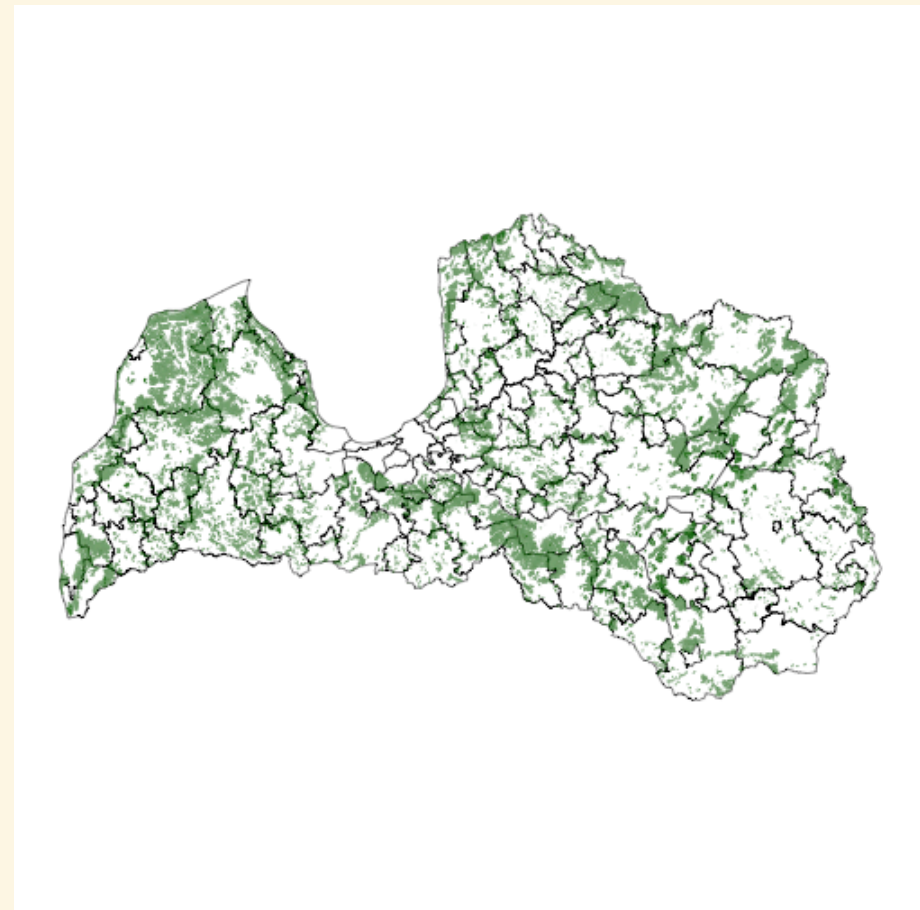
- Maps - import of shapefiles, simplefeatures, analysis and plotting

```
library(rgdal)
library(ggplot2)
novadi <- readOGR(dsn="GIS_Latvija10.2.gdb",
                 layer="novadi_2009_pol")

mezi <- readOGR(dsn="lvm_kvartali_shp",
               layer = "LVM_KVARTALI")

novadi.df<-fortify(novadi,region="Nos_pilns")
mezi.df <- fortify(mezi, region = "BLOCKKEY")

ggplot(novadi.df, aes(long,lat, group = group)) +
  geom_polygon(fill = "white", color = "black",
              size = 0.25) +
  geom_polygon(data = mezi.df, fill = "darkgreen",
              alpha = 0.9) +
  coord_fixed() + theme_void()
```



Functions and packages

- Routine calculations can be made as functions
- Many functions - new R package
- Data can be distributed as package

```
my_function <- function(data, k = 2) {  
  (data * k) ^ k  
}  
  
my_function(5)
```

```
## [1] 100
```

Reproducible research and dynamic documents

- Git integration
- Markdown language with R code integration
- R libraries `rmarkdown`, `bookdown`, `blogdown`

What next?

What you want to learn?

- R basics?
- Data visualisation?
- Data importing from different sources?
- Data filtering, selecting, summarising?
- Statistical analysis?
- Other?

- Thank you for the attention!



bit.ly/WaterAct-project



bit.ly/WaterAct-Researchgate



WaterAct
Joint actions for more efficient management
of common groundwater resources