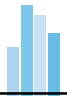
Introduction of R at Stats Denmark – a not entirely completed how-to

Presentation at uRos2018

Peter Tibert Stoltze
Head of Methods and Analysis
Statistics Denmark

STATISTICS DENMARK



Main points

- We want to use R for a number of reasons
- We already use R for a few specific tasks
- We have a pretty cool R environment
- We have made a (draft) strategy for introducing R, and the strategy goes well beyond introducing R

We want to use R because…

- R is widely used in academia
- New staff knows R
- Many useful packages specifically for official statistics
- Modern tool easy to integrate with other software
- Presumably low cost (not that important but still nice…)
- R was sneaking in anyway
- R is very cool
- Add your own cause…

STATISTICS DENMARK 3

Main points

- We want to use R for a number of reasons
- We already use R for a few specific tasks
- We have a pretty cool R environment
- We have made a (draft) strategy for introducing R, and the strategy goes well beyond introducing R

STATISTICS DENMARK 4

STATISTICS DENMARK

We already use R for…

- Designing certain survey samples at the methodology unit
  - E.g. **`stratification`** package to decide size classes within NACE groups rather than just applying a common cut
  - Obviously, following yesterdays workshop, we will start using **`SamplingStrata`** instead (thanks Giulio!)
- Standardized and modernized data editing (SMOF)
  - **`validate`** package and friends following templates
  - Execution embedded in a GUI such that no R knowledge is demanded from the user…

STATISTICS DENMARK 5

SMOF steps

- Steps performed by Methodology staff
  - Make a script (using a template and knowledge from domain experts)
  - Setup script for execution in GUI
- Steps performed by domain experts
  - Execute script from GUI
  - Check log after execution
  - Inspect editing report
  - Check individual observations
    - Different kind of actions…

STATISTICS DENMARK 6

STATISTICS DENMARK

Main points

- We want to use R for a number of reasons
- We already use R for a few specific tasks
- We have a pretty cool R environment
- We have made a (draft) strategy for introducing R, and the strategy goes well beyond introducing R

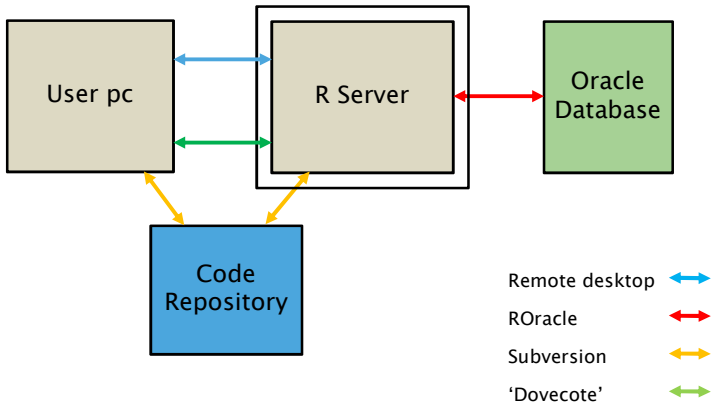
STATISTICS DENMARK 7

Tech info

- Central installation
  - Rstudio Server Pro
  - Secure, all access to data is logged
- Server specifications
  - Windows Server 2012 R2
  - 20 Intel Xeon E5-2660 v3 cores @ 2.6 GHz
  - 384 GB RAM
- Package management
  - Based on Packrat

STATISTICS DENMARK 8

R runs in a shielded environment



STATISTICS DENMARK 9

Main points

- We want to use R for a number of reasons
- We already use R for a few specific tasks
- We have a pretty cool R environment
- We have made a (draft) strategy for introducing R, and the strategy goes well beyond introducing R

STATISTICS DENMARK 10

STATISTICS DENMARK

Ambitions

- We don't want to make the same mistake as when desktop SAS was introduced 20 years ago
  - Not meant as SAS bashing, but the result of virtually no rules is not pretty ☹
- We want to promote a professional approach to writing and working with computer code and data
  - Code shall live in a CVS (Concurrent Versions System)
  - Data shall live in data bases (and not in files on various network drives)

STATISTICS DENMARK 11

Code maturity

- We have outlined three levels of code maturity:
  1. Ad hoc code
  2. Code residing in a CVS (Subversion) and using data from a data base (Oracle)
  3. As 2. but reviewed by a peer (who qualifies as peers remains to be decided)
- The goal is to make explicit decisions at management level regarding what level of code maturity is needed

STATISTICS DENMARK 12

Enforce semi-rigid GSBPM compliance

- Production process aligned with GSBPM
  - Process documentation as per GSBPM
  - Folder structure as per GSBPM
  - Scripts directed toward specific processes as per GSBPM
- A number of principal advantages
  - Reuse code/methods across domains
  - Rotating staff internally
  - Follow progress towards dissemination (dashboard style management information)
  - Methodology staff do not have to waste their time finding the appropriate point in programs and worry about side effects…

STATISTICS DENMARK 13

Enforce semi-rigid GSBPM compliance

- Our working hypothesis is that introducing what seems like a rigid framework actually allows for quicker adaptation of changes
  - The expected input and the generated output are both well defined, hence side effects should be minor
- It also allows for appropriate choice of tools for each process, e.g. Python rather than R or even reuse of existing SAS-code
  - Sometimes you need more muscles than brain…

STATISTICS DENMARK 14

STATISTICS DENMARK

Analysis Tool	Similar Superhero	Super Powers in Common
R 	Batman 	• Detective Work • Intelligence • Cunning • Usage of Tools • More Brain than Muscles
Python 	Superman 	• Muscle Power • Super Strength • Elegance • Wide Range • More Muscles than Brain

STATISTICS DENMARK 15

Realizations

- If introduction of R seems difficult, then R is probably not to blame…
- Introduction of new technologies will happen more often in the future (i.e. tomorrow), so we might as well start practicing!

STATISTICS DENMARK 16

STATISTICS DENMARK

Main points

- We want to use R for a number of reasons
- We already use R for a few specific tasks
- We have a pretty cool R environment
- We have made a (draft) strategy for introducing R, and the strategy goes well beyond introducing R

STATISTICS DENMARK 17

Thanks for the attention - and thanks to the organizers!

Email: psl@dst.dk
Twitter: @MetodePeter

STATISTICS DENMARK 18