

# Free and Open Source Software - R

## in our statistical office – why?

Alexander Kowarik  
Methods Unit

Pisa, 20.9.2022

[www.statistik.at](http://www.statistik.at)

Independent statistics for evidence-based decision making

# Outline

- Computing in the Statistical Production
- Learn from DevOps
- Open Source Tool Development



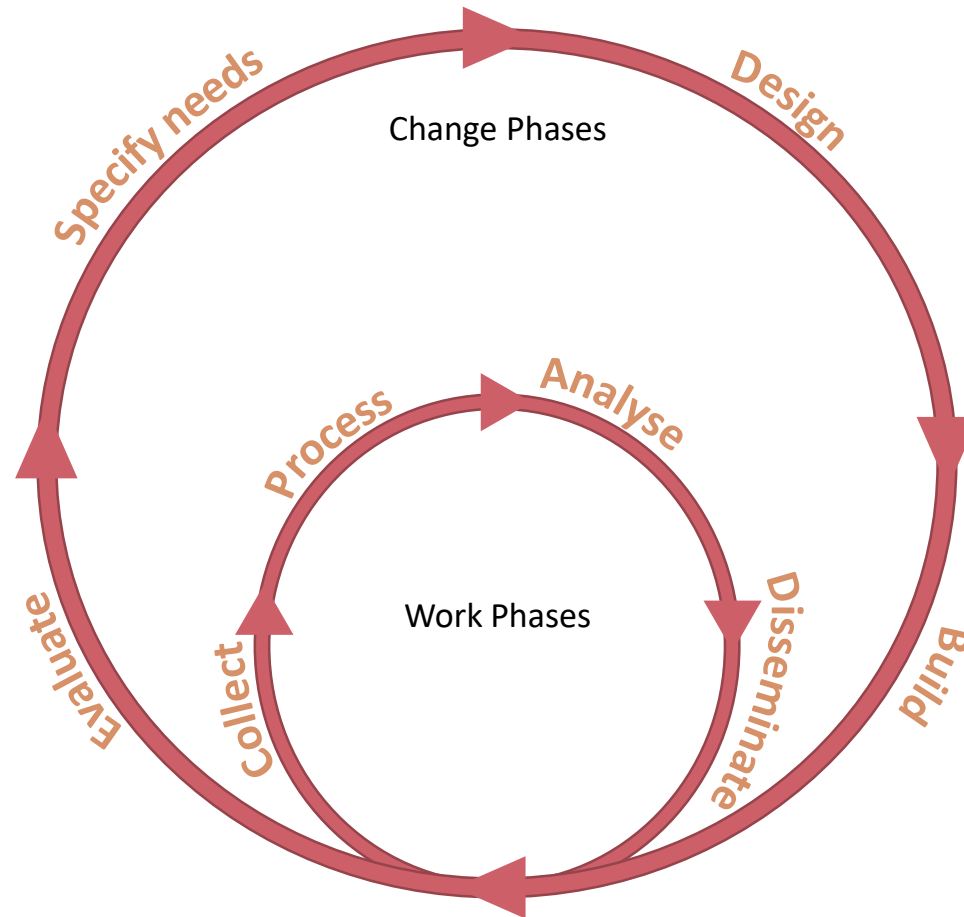
## About myself:

- Master and PhD in technical mathematics with focus on computational statistics
- Since 2008 in the methods unit @Statistics Austria
- Since 2014 head of the methods unit
- (Co-) author of several R packages

# Computing in the statistical office II

- Tasks are split up between IT, methodologist and subject matter expert
  - IT Development skills
  - Mathematical, numerical and statistical skills
  - Expert knowledge in the field
- A task similar to Research Software Engineering: <https://society-rse.org>

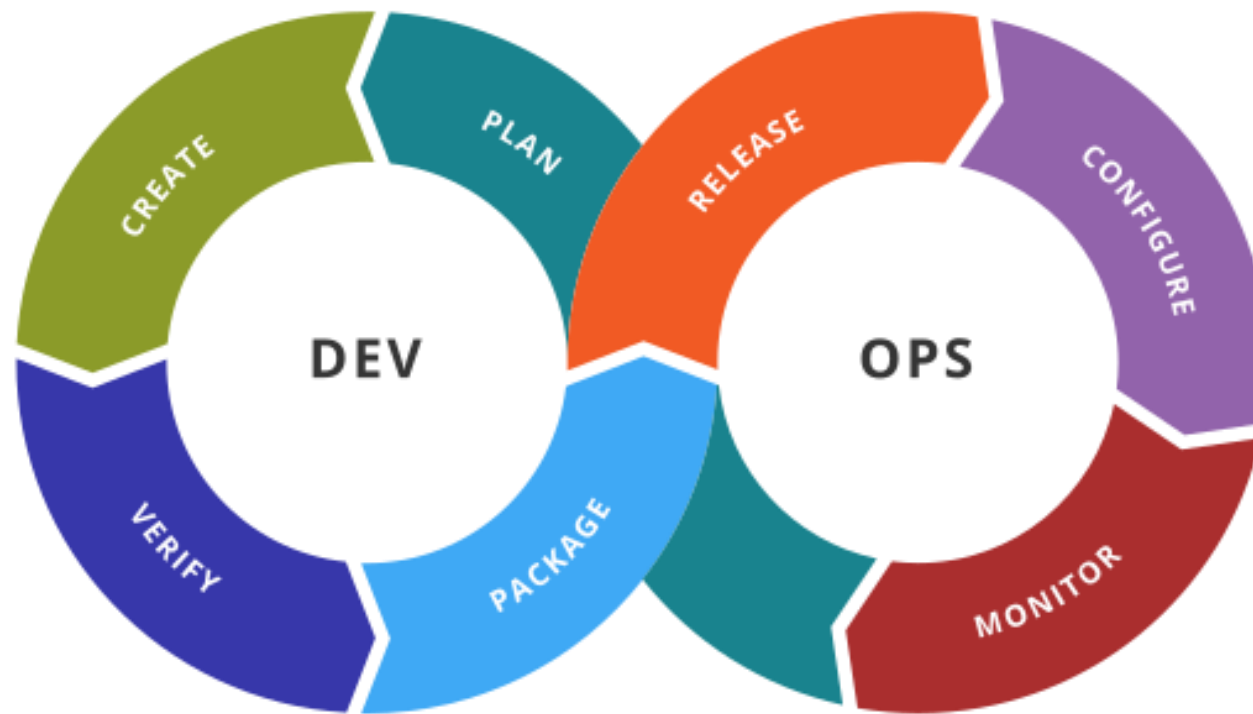
# Statistical Production II



In GSBPM, there are some phases which are undertaken quickly and frequently – the Work Phases.

There are other phases which are undertaken less often - the Change Phases.

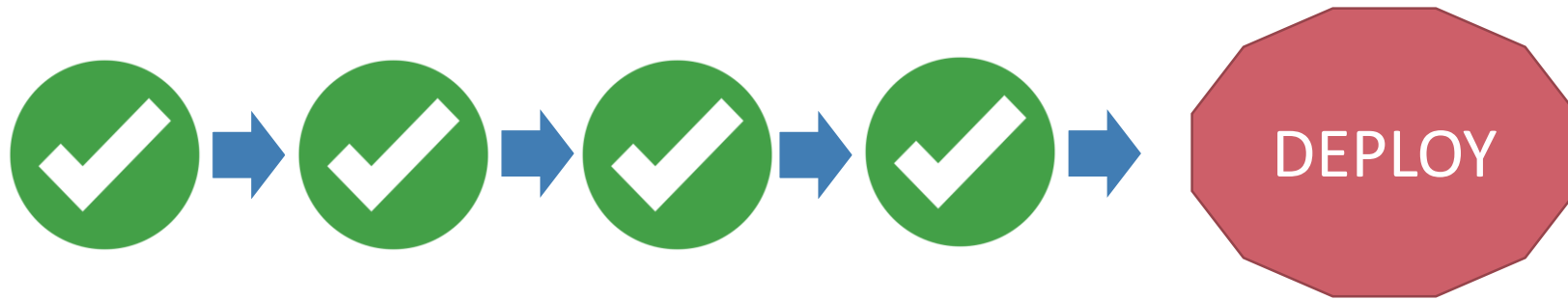
<https://statswiki.unece.org/display/GSBPM/Clickable+GSBPM+v5.1>



Kharnagy, CC BY-SA 4.0 <<https://creativecommons.org/licenses/by-sa/4.0/>>, via Wikimedia Commons

- Learning from DevOps (development and operations) practices in system development
- A few ideas can be “translated” to the statistical production process

# Automate Everything (as much as possible)



- Continuous Integration and Continuous Delivery (CI/CD) pipelines for data processing
- Quickly react to updated data or improved methods
- Data science, machine learning and statistical methods as automated steps in a streamlined process
- Infrastructure as code

# Testing / Observing Quality Dimensions

- Unit testing of software components
- Plausibility checks for data (automated data editing)
- Quality controls/indicators throughout the process including traditional statistical quality dimensions
- Trigger manual intervention in case of „extreme quality events“



# Foster Continuous Improvement

- Modularity allows to quickly integrate new methods, new data sources or new software tools.
- Integrate teams: statisticians, subject matter experts and developers



<https://freshideen.com/trends/lego-spiele-im-begriff-einen-vorbildlichen-schritt-zu-tun.html>



# Be Reproducible

- Making the data generation process transparent
- Data versioning
- GUIs for data editing -> produce code

Image: <https://www.flickr.com/photos/tjeerd/34733471821>

# What tools do you need?

- A code based statistical software
- Version control -> GIT
- Pipeline tool, e.g. Gitlab CI, Github Actions, Jenkins
- Storage to deploy artifacts, e.g. R packages to a CRAN-like mirror, web applications to a server, etc.

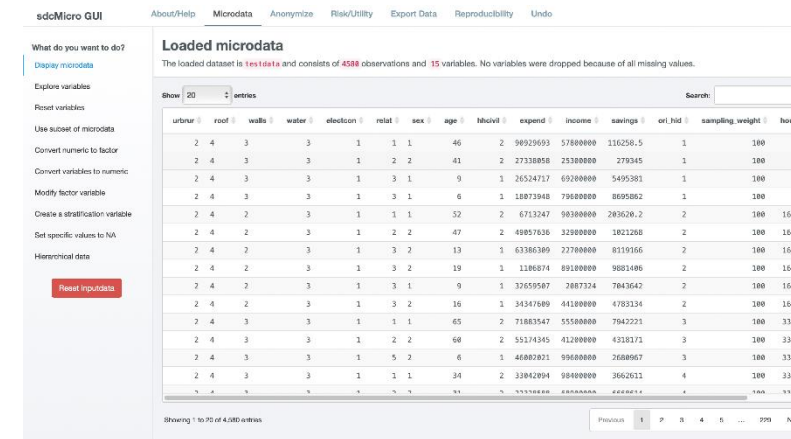
# Collaborative developments

## ESSNet projects to develop/maintain SDC software

- **muArgus + tau Argus:** Initial development by CBS et al as closed software
- Later published as open source
- R packages **sdcMicro + sdcTable** developed initially in parallel
- Convergence of the developed methods
  
- **Now:** newly implemented method „Target Record Swapping“ as OS C++ library used by sdcMicro and muArgus
  
- Contributors from outside the project, worldwide usage of tools

# 6.4 Apply disclosure control

- sdcMicro: <https://sdctools.github.io/sdcMicro/>
- Disclosure risk estimation
- Microaggregation, (correlated) noise , local suppression, ...
- Shiny application with
  - reproducible output

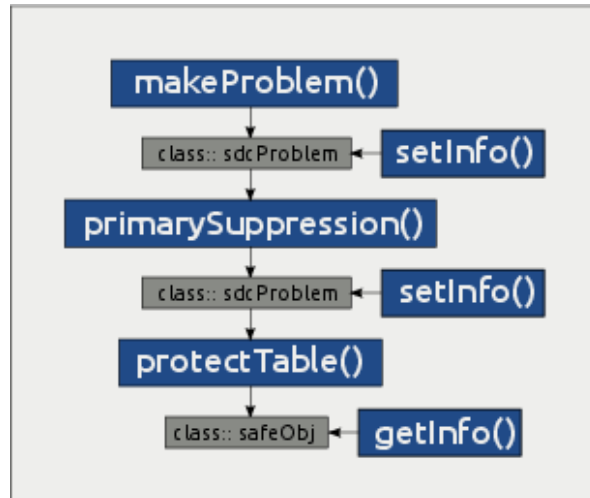


The screenshot shows the sdcMicro GUI interface. The main panel displays a table of loaded microdata with the following columns: urban, roof, walls, water, electoon, relat, sex, age, hbvvl, expend, income, savings, crl\_hid, sampling\_weight, and household. The table contains 4588 rows of data. The interface includes a sidebar with navigation options like 'What do you want to do?', 'Explore variables', and 'Reset variables'. The top navigation bar includes 'About/Help', 'Microdata', 'Anonymize', 'Risk/Utility', 'Export Data', 'Reproducibility', and 'Undo'.

urban	roof	walls	water	electoon	relat	sex	age	hbvvl	expend	income	savings	crl_hid	sampling_weight	househd
2	4	3	3	1	1	1	46	2	96929693	57800000	116258.5	1	100	
2	4	3	3	1	2	2	41	2	27338856	25300000	278345	1	100	
2	4	3	3	1	3	1	9	1	26524717	69200000	549581	1	100	
2	4	3	3	1	3	1	6	1	18073048	79600000	869582	1	100	
2	4	2	3	1	1	1	52	2	6713247	98300000	283628.2	2	100	16.6
2	4	2	3	1	2	2	47	2	49857636	32900000	1821208	2	100	16.6
2	4	2	3	1	3	2	13	1	63386389	22700000	8119166	2	100	16.6
2	4	2	3	1	3	2	19	1	1186874	89100000	9881486	2	100	16.6
2	4	2	3	1	3	1	9	1	32659507	2887324	7843642	2	100	16.6
2	4	2	3	1	3	2	16	1	34347689	41100000	4783334	2	100	16.6
2	4	3	3	1	1	1	65	2	71883547	55500000	7842221	3	100	33.3
2	4	3	3	1	2	2	68	2	55174345	41200000	6318171	3	100	33.3
2	4	3	3	1	5	2	6	1	46802021	99000000	2680967	3	100	33.3
2	4	3	3	1	1	1	34	2	33842894	98800000	3662611	4	100	33.3

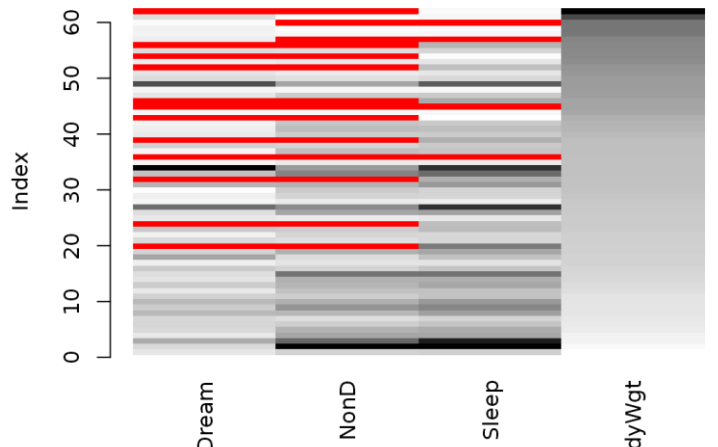
## 6.4 Apply disclosure control I

- sdcTable: <https://sdctools.github.io/sdcTable/>

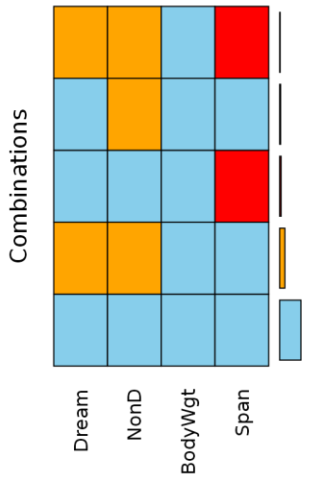
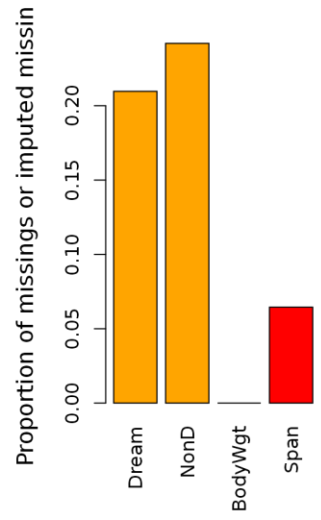
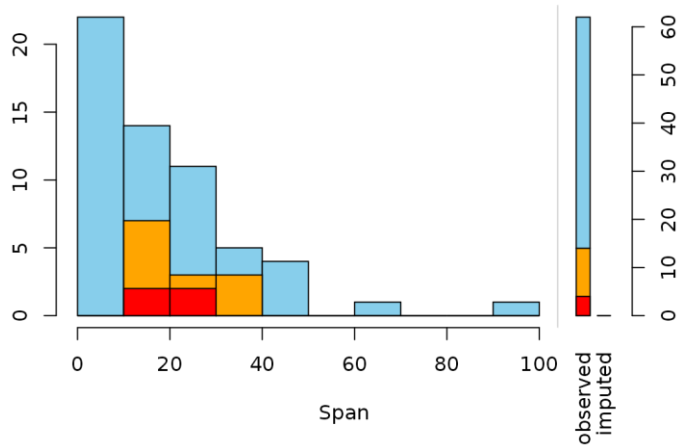


W	A	B	C	Total
X	20	50	10	80
Y	S	19	P	49
Z	S	32	S	61
Total	45	101	44	190

# 5.4 Edit and impute (6.2 validate outputs)



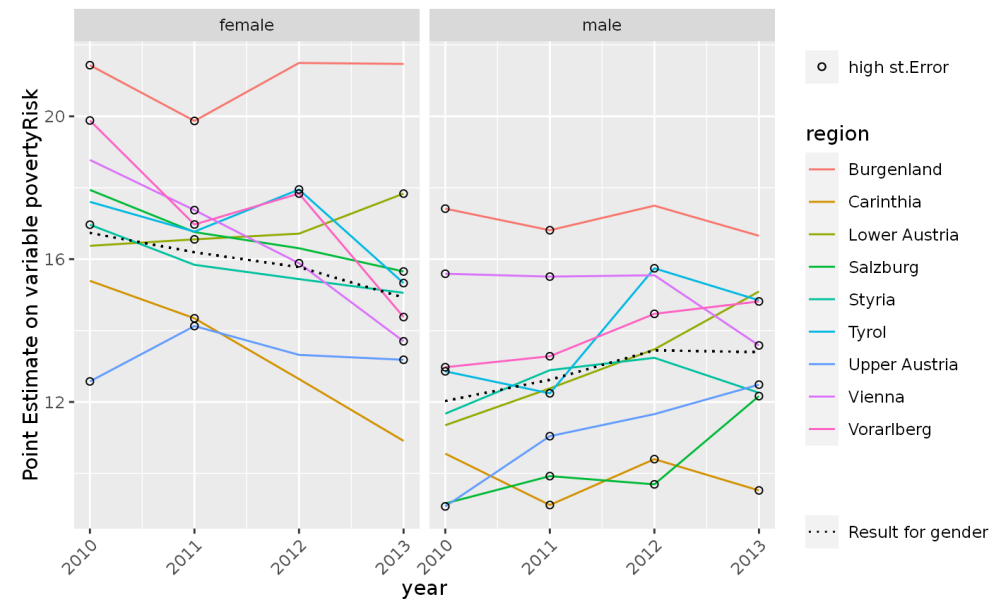
- R package VIM:
- Visualization and Imputation of Missing Values
  - Imputation:
    - Donor-based: kNN, hotdeck, matchImpute
    - Model-based: irmi, regressionImp



## 5.6 Calculate weights (+ analyze 6.1 - 6.3)



- Calculate weights - calibration
  - Draw bootstrap samples with calibration
  - Estimate errors and confidence intervals
  - Some simple visualizations to analysis results
- 
- <https://statistikat.github.io/surveysd>



# Open Source Tools – Use, Contribute, Modify

## What is the awesome list?

Awesome list of software for official statistics



awesome

www.awesomeofficialstatistics.org

Search or jump to... Pulls Issues Marketplace Explore

SNStatComp / awesome-official-statistics-software

Unwatch 30 Unstar 161 Fork 41

Code Issues Pull requests 1 Actions Projects Wiki Security Insights

Social interactions

### Awesome official statistics software

An awesome list of open source statistical software packages useful for creating and accessing official statistics.

#### Criteria

An item on this list is awesome because

1. it is free, open source, and available for download and
2. it is confirmed to be used in the production of official statistics by at least one institute or it provides access to official statistics publications.

We prefer packages that are easy to install and use, have at least one stable version, and are actively maintained. Contributions are welcome.

### License



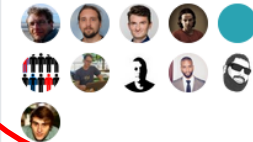
This work is licensed under a Creative Commons Attribution 4.0 International License.

#### Open license

An awesome list of statistical software for creating and accessing official statistics

official-statistics gsbpm

#### Contributors 15



+ 4 contributors

#### Working together

#### Contributions

Awesome contributions are welcome, here are ways to do it:

- The GitHub way: send us a pull request to add directly to this list.
- Add an item to the issue tracker issue tracker. (you need a GH account)
- Send an e-mail to mark dot vanderloo at gmail dot com or olav dot tenenbach at gmail dot com or tweet @markvdloo