Why and for what purpose R in Official Statistics?

- uRos 2024 -

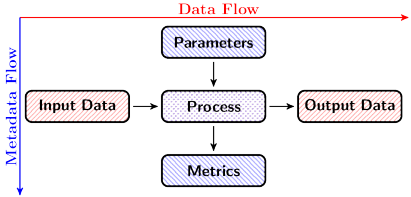
Sandra Barragán

- Statistics Spain (INE) -

Instituto Nacional de Estadística

Conclusions: Take home messages

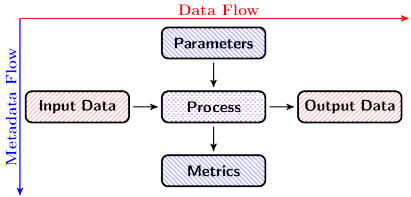
- R is open source, user friendly (easy to learn), constantly evolving. R has a bunch of packages implementing statistical methods and a wide community of users.
- R facilitates a proper **implementation** and **automation** of **modular processes** in the production of Official Statistics.



Conclusions: Take home messages

- R is open source, user friendly (easy to learn), constantly evolving. R has a bunch of packages implementing statistical methods and a wide community of users.
- R facilitates [illegible] **n** and **automation** of **modular processes** in the production of Official Statistics.

HOW?



INE

Outline

- Introduction
- Industrialization of Official Statistical production
- Implementation
- Use Cases implemented in R
- Conclusions

INE

Introduction

Run, or he's going to tell us about **R** again!

R is new...

Some people are afraid of R language because it is something new (for them).

NEW???

... not so new...

1993-1999

R was created by Ross Ihaka and Robert Gentleman, statisticians from the University of Auckland, New Zealand.


Highlights

1995 GNU License

1996 "R: A Language for Data Analysis and Graphics" Journal of Computational and Graphical Statistics

1997 CRAN is created by Kurt Hornik and Friedrich Leisch

2000-2009

The first stable version of R, version 1.0.0, is released on the 29th February.

Highlights

2001 R News, Bioconductor

2004 R!

2005 CRAN Task Views

2007 ggplot2

2010-2019

RQpenSci: "Transforming science through open data, software & reproducibility".

The future of interactive graphics in R
Hadley Wickham

Highlights

2011 R Studio

2012 R, Shiny

2014 rmarkdown

2017 CRAN surpasses **10,000** packages.

2020-2024

R.4.0.0 is released
CRAN > 15000 packages

Highlights

2021

2022 posit

2023 CRAN surpasses 2 **0,000** packages.

2024 R packages official DOI

2010

CRAN Task View: Official Statistics & Survey Statistics is commited for the first time.

Highlights

Maintained by: Matthias Templ, Alexander Kowarik, Tobias Schoch

Part 1: "Producing Of. Stat." Part 2: "Access to Of. Stat." Part 3: "Related methods"

>131 packages

2013

First uRos (use of R in Official Statistics) is held in Romania.

uRosCont

Highlights

2014

First International edition of uRos

2017

The list of Awesome official statistics software started during the UNECE data editing conference (SDE 2017)

awesome

Highlights

Maintained by: Olav ten Bosch and other 24 contributors

Following the GSBPM

135 packages (>70% R packages)

2020-2024

Increment of the use of R in the Statistical Offices (including in production processes)

Highlights

2023 NTTS23: session R, Python, Julia

2024 ESTP course: R, Python and Julia: do you know them all?

(Source: Jumpingrivers)

R in official statistical production

- **CRAN Task View: Official Statistics & Survey Statistics**
  https://cran.r-project.org/web/views/OfficialStatistics.html

CRAN Task View: Official Statistics & Survey Statistics

Maintainer:	Matthias Templ, Alexander Kowarik, Tobias Schoch
Contact:	matthias.templ at gmail.com
Version:	2023-02-19
URL:	https://CRAN.R-project.org/view=OfficialStatistics
Source:	https://github.com/cran-task-views/OfficialStatistics/
Contributions:	Suggestions and improvements for this task view are very welcome and can be made through issues or pull requests on GitHub or via e-mail to tl guide.
Citation:	Matthias Templ, Alexander Kowarik, Tobias Schoch (2023). CRAN Task View: Official Statistics & Survey Statistics. Version 2023-02-19. URL
Installation:	The packages from this task view can be installed automatically using the ctv package. For example, `ctv::install.views("OfficialStatistics"`. `ctv::update.views("OfficialStatistics")` installs all packages that are not yet installed and up-to-date. See the CRAN Task View Initiative for

This CRAN Task View contains a list of packages with methods typically used in official statistics and survey statistics. Many packages provide functions for mo strict categorization and packages may be listed more than once.

The task view is split into several parts

- First part: "Producing Official Statistics". This first part is targeted at people working at national statistical institutes, national banks, international organizat statistics and using methods from survey statistics. It is loosely aligned to the "Generic Statistical Business Process Model".
- Second part: "Access to Official Statistics". This second part's target audience is everyone interested to use official statistics results directly from within R.
- Third part: "Related Methods" shows packages that are important in official and survey statistics, but do not directly fit into the production of official statist collection of packages that are loosely linked to official statistics or that provide limited complements to official statistics and survey methods.

First Part: Production of Official Statistics

1 Preparations/ Management/ Planning (questionnaire design, etc.)

- questionr package contains a set of functions to make the processing and analysis of surveys easier. It provides interactive shiny apps and addins for data re several convenience functions.

R in official statistical production

- **Awesome official statistics software**
  https://github.com/SNStatComp/awesome-official-statistics-software

Awesome official statistics software awesome

An awesome list of open source software for official statistics.

An item on this list is awesome because it is

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

We prefer software that is easy to install and use, has at least one stable version, and is actively maintained.

Design frame and sample (GSBPM 2.1)

- CRAN 1.5-4 – 5 months ago license GPL (>= 2)
  R package SamplingStrata. Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys.
- CRAN 1.0.4 – a year ago license EUPL
  R package R2BEAT. Multistage Sampling Allocation and PSU Selection.

Design variable descriptions (GSBPM 2.2)

INE

Context

- **Eurostat: Working group on Open Source Software for the use in Statistics**
- **European Comission: Open source software strategy**
  https://commission.europa.eu/about-european-commission/departments-and-executive-agencies/informatics/open-source-software-strategy_en
- **European Comission: Open Source Observatory (OSOR)**
  https://joinup.ec.europa.eu/collection/open-source-observatory-osor

INE


OPEN SOURCE
OFFICIAL STATISTICS
SAS
imgflip.com

From SAS to R...

The 12th International Conference Use of R in Official Statistics

uRos2024

27-29 November 2024, Piraeus, Greece

November 29, Friday

8^{30} – 9^{00} **Walk In**

9^{00} – 10^{00} ***Keynote Speaker 2:*** Romain Lesur, INSEE France

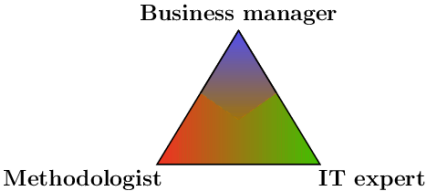
Presentation - Moving away from SAS: an opportunity to modernise the practices of statisticians

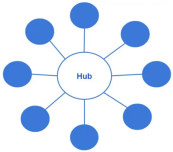
Industrialization of Official Statistical production

INē

Industrialization of official statistical production

- Standardization
- Automation
- Modularity





Industrialization of official statistical production

- Standardization
- Automation
- Modularity

Advantages:

+ Reuse of processes.
+ Easier to detect bugs and errors.
+ Maintenance.
+ Resource optimisation.
+ Evolution.

Industrialization of official statistical production: Standardization

Focused on the process instead of the product, but flexible enough to be adapted to the particularities of each product.

- **Parameterisation:** Same process can be executed for different products without changes in the code, just changing the parameters.
- **Reproducibility:** consistent execution when it is repeated.
- **Documentation:** every part of the process is documented.
- **Quality** control.
- Continuous integration and deployment: **CI/CD**.

See paper in JOS: *Data organisation and process design based on functional modularity for a standard production process*, Salgado et al. [2018].

INE

Industrialization of official statistical production: Automation

Implementing manual processes to be executed by a machine.
Scheduling task to be executed based on a calendar.

Ex1: Generate files from collection to a repository at a time everyday.
Ex2: Trigger the execution of selective editing once all the files are available and a date is reached.

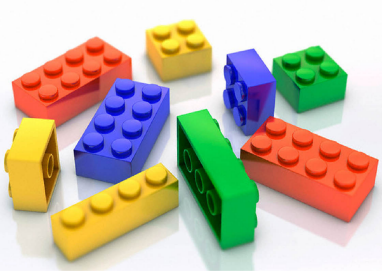
Industrialization of official statistical production: Automation in R

Scheduling R processes in the system from R directly:

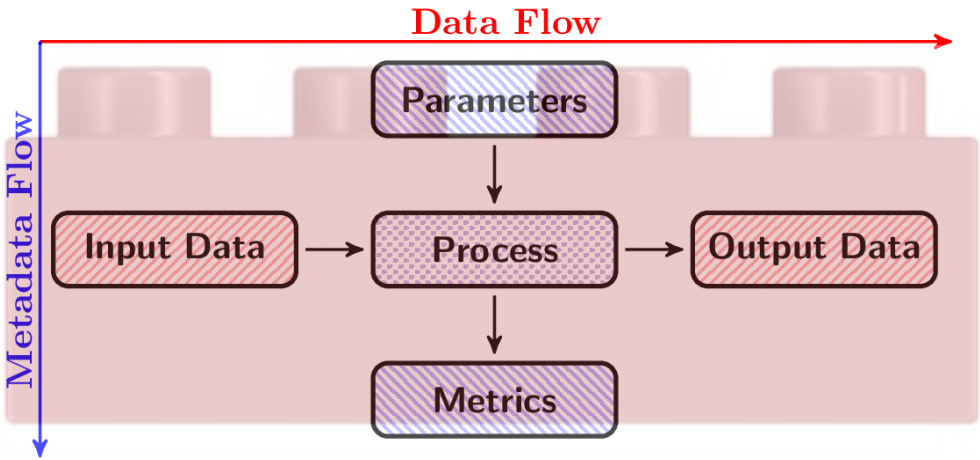
- `cronR` in Linux:
  - Create a job: `cron_add()`.
  - Remove a job: `cron_rm()`.
  - List all jobs scheduled: `cron_ls()`.
- `taskscheduleR` in Windows:
  - Create a job: `taskscheduler_create()`.
  - Remove a job: `taskscheduler_delete()`.
  - List all jobs scheduled: `taskscheduler_ls()`.

Different from the Background jobs of RStudio:
https://docs.posit.co/ide/user/ide/guide/tools/jobs.html.

Industrialization of official statistical production: Modularity

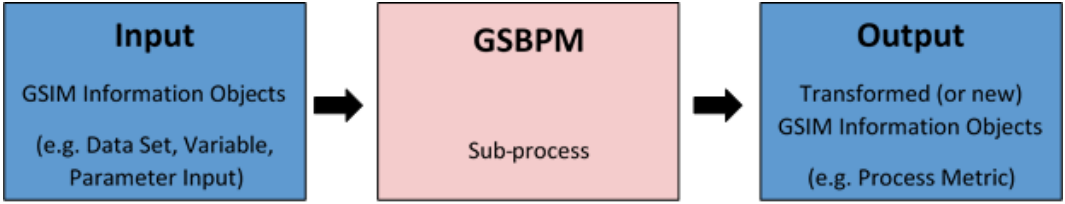


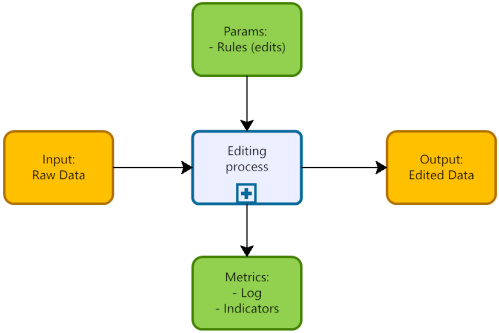
Industrialization of official statistical production: Modularity



INE

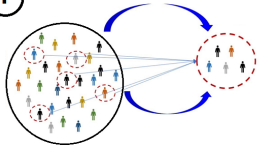
Industrialization of official statistics: Modularity





Industrialization of official statistics: Modularity

1



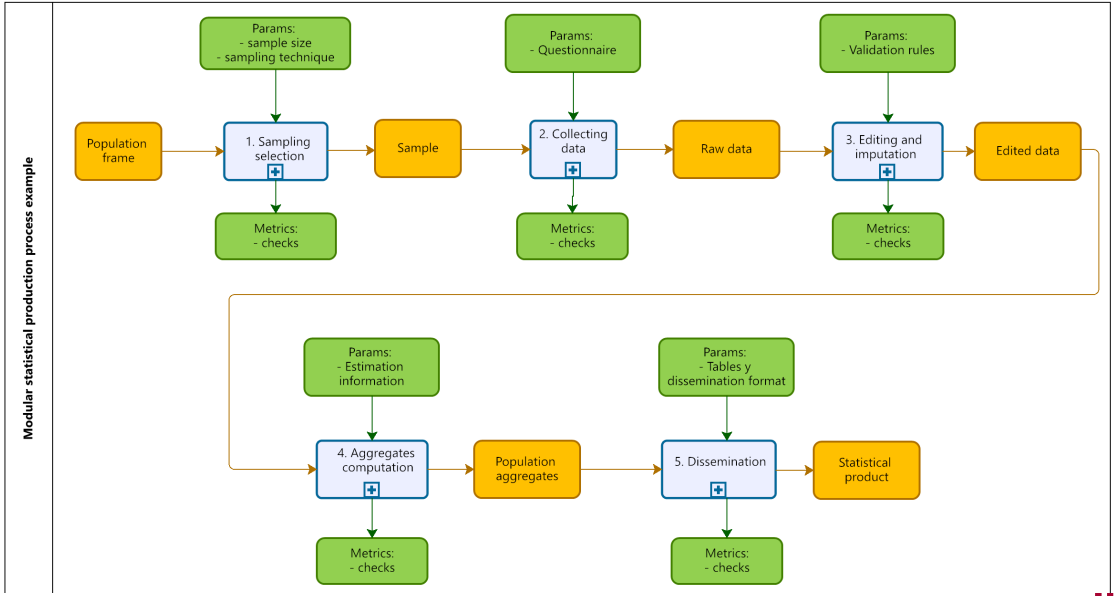
2

3

4

5

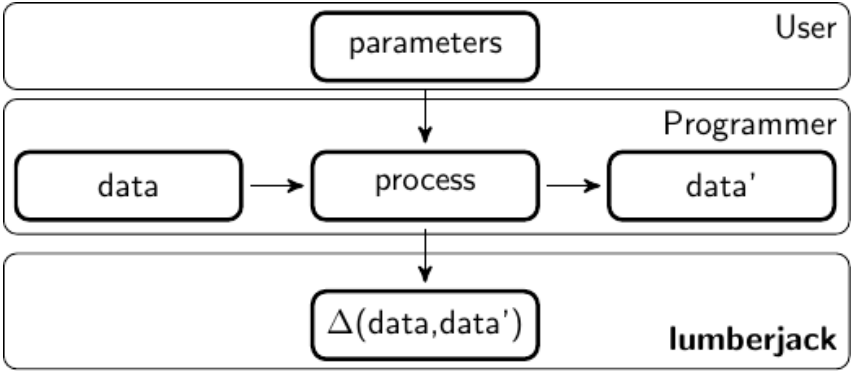
Industrialization of official statistical production: Modularity



Modularity in R

Tip!

R package `lumberjack`: https://github.com/markvanderloo/lumberjack



Implementation

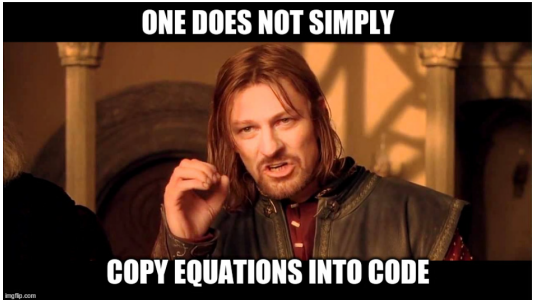
Implementation

If you can't describe what you are doing as a process,
you don't know what you're doing.

W. Edwards Deming

Implementation

Implementing methods is not trivial. See van der Loo [2020].



Principles[1]:

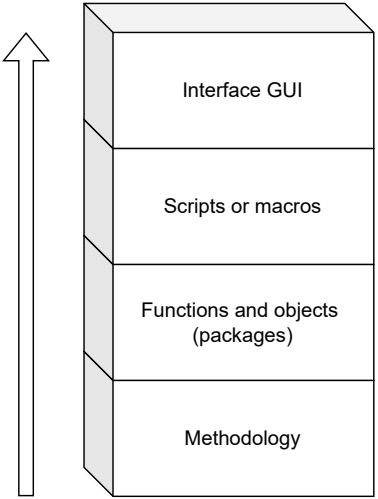
- Funcionality
- Object-oriented

R[2] can help you.

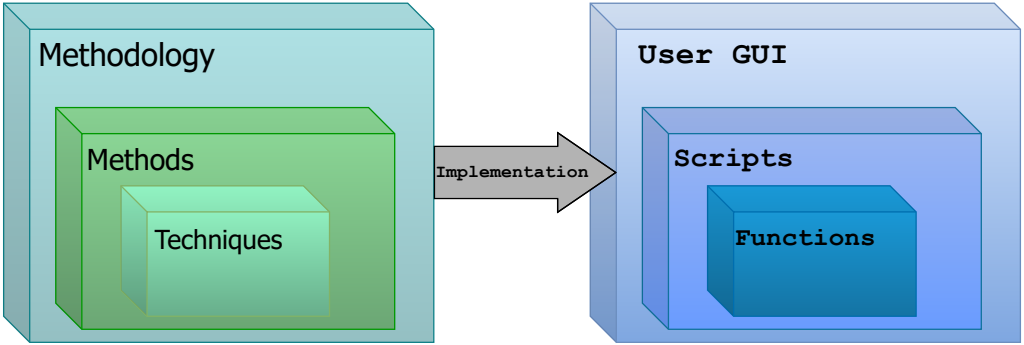
[1] See Tucker and Noonan [2007]
[2] See Matloff [2011]

Implementation

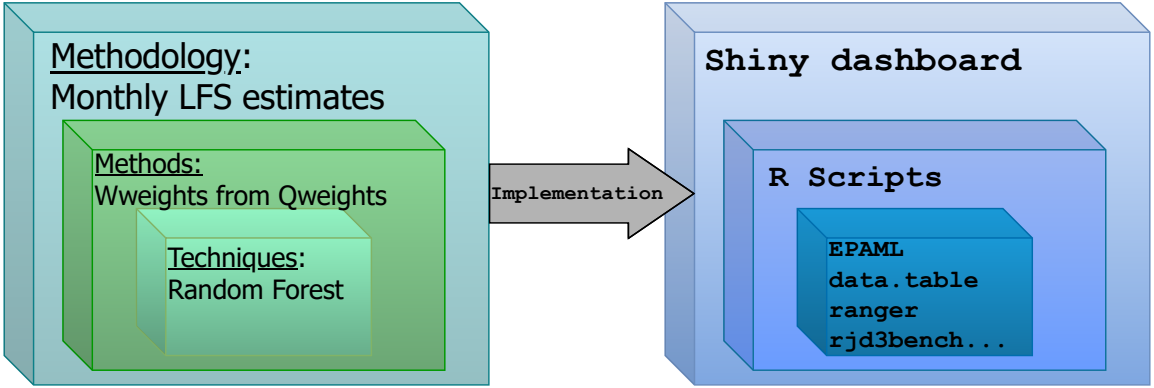
Traditional levels of implementation



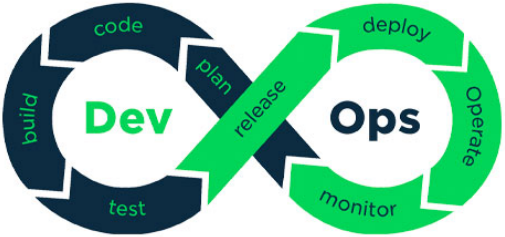
Implementation

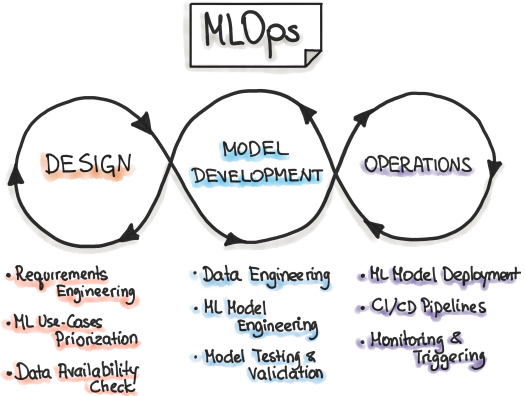


Implementation: example

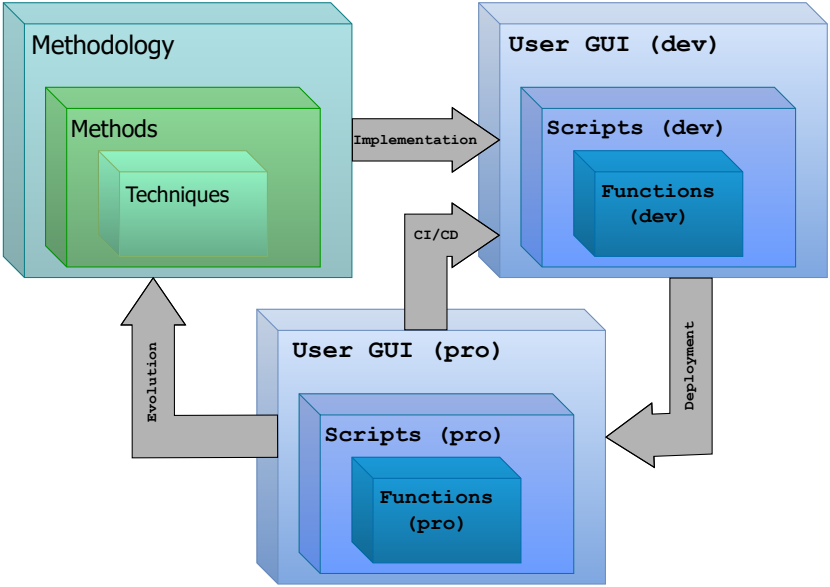


Implementation for production: DevOps and MLOps





Implementation for production: StatOps



R in production

- **P**roduction vs. **p**roduction
- Project oriented (usethis, here)
- Environments: production, test, development.

More details in: R in production - Hadley Wickham

INE

R in production

Relevant considerations:

1. Running code on another machine
   - Someone else. R package `pointblank`.
   - Debugging, Logging
   - Configuration (Authentication, package installation)
2. Running code multiple times (Changes over time)
   - Data schema
   - Package versions: R package `pak`.
   - System libraries
   - Configuration (OS, Requirements...)
3. Shared responsability
   - Parquet format: R packages `arrow` and `nanoparquet`.
   - Code review: R packages `cyclocomp` and `covr`.
   - Git

INE

Implementation: Some good practices

- Version control: Git (github, gitlab)
- Code organization:
  - Naming: variables names.
  - Indentation: margins, spaces.
  - Well structured documentation (comments).
  - Functions and modularity.
  - Check arguments validity.
- Files and folders:
  - Naming: files names.
  - Well structured organization.
- Good habits:
  - Release memory.
  - Executing from the beginning.

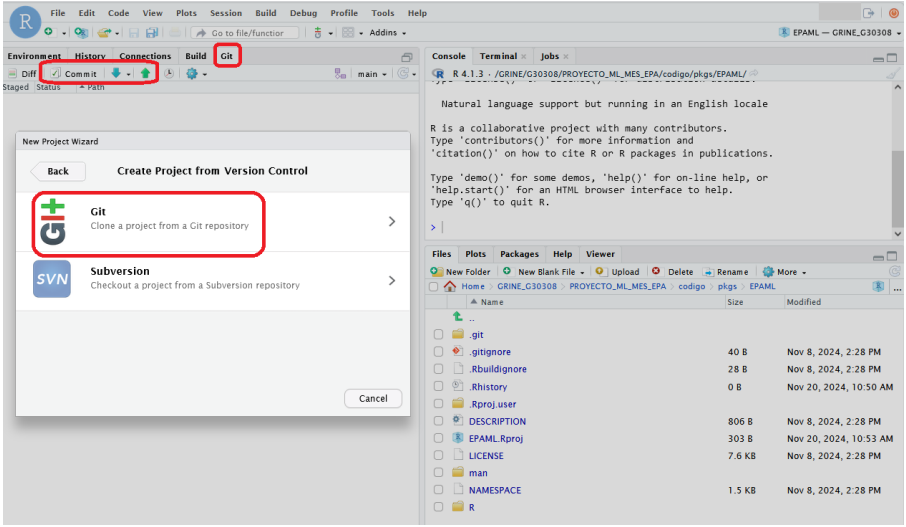
RStudio eases your work

```
#####                          ::::::::::::::::                          ####
#####                                PROCESO                             ####
## * Leer microdatos                                                      ####
schema_raw.stSchema <- StxlsxToSchema(
  xlsxName    = file.path(path_param, schema_raw_filename),
  sheetToRead = schema_raw_sheetname
)

microdata_raw.dt  <- fread_fwf(
  filename    = file.path(path_data, data_raw_filename),
  StfwfSchema = schema_raw.stSchema,
  outFormat   = "data.table", perl = TRUE,
  validate    = validate_data, convert = convert_types
)

microdata_recontacto.dt  <- fread_fwf(
  filename    = file.path(path_data, data_recontacto_filename),
  StfwfSchema = schema_raw.stSchema,
  outFormat   = "data.table", perl = TRUE,
  validate    = validate_data, convert = convert_types
)

## * Leer estructura de la cnae09                                         ####
cnae09.dt <- as.data.table(read.xlsx(
  xlsxFile = file.path(path_param, cnae09_estructura_filename),
  sheet    = cnae09_estructura_sheetname))

cnae09_clases.dt <- cnae09.dt[nchar(COD_CNAE2009) == 4]
setnames(cnae09_clases.dt, 'COD_CNAE2009', 'cnae09_clase')
cnae09_grupos.dt <- cnae09.dt[nchar(COD_CNAE2009) == 3]
setnames(cnae09_grupos.dt, 'COD_CNAE2009', 'cnae09_grupo')
cnae09_divisiones.dt <- cnae09.dt[nchar(COD_CNAE2009) == 2]
setnames(cnae09_divisiones.dt, 'COD_CNAE2009', 'cnae09_division')
```

```
::::::::::::::
PAQUETES
::::::::::::::
RUTAS
::::::::::::::
FUNCIONES
::::::::::::::
PARAMETROS
* Parámetros para lectura
* Parámetros generales
* Parámetros para depuración DOS
* Parámetros para depuración selectiva
* Parámetros para depuración automática
::::::::::::::
PROCESO
* Leer microdatos
* Leer estructura de la cnae09
* Generar variables derivadas
:=============:
DOS
* Leer edits
* Aplicar edits y tratar errores
** sample-wise
** record-wise
*** norden_unico
*** clase_longitud
*** (variable)_formato
*** (variable)_recorrido
*** cifraNep(, Nac, Ext)_NA -- balance
*** ocupados(, Fijos, Event)_NA -- balance
** imputación NA
:=============:
Depuración selectiva
:=============:
Depuración interactiva
:=============:
Depuración automática
** Localización de errores
** Reemplazamiento de errores por NA
** Imputar NA en edits de balance activos
:=============:
```

INE

RStudio eases your work



Implementation: Some good practices

More details in:

- R Style Guide de Google.
- Advanced R by Hadley Wickham (Wickham [2019]).
- Writing Better R Code workshop.
- Boswell and Foucher [2011].
- R packages `goodpractice`.

Not just the code...

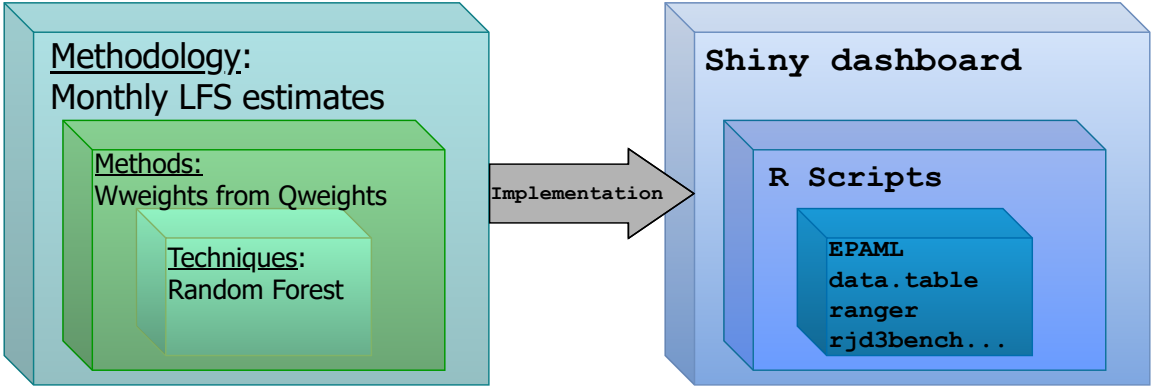
Use Cases implemented in R

INē

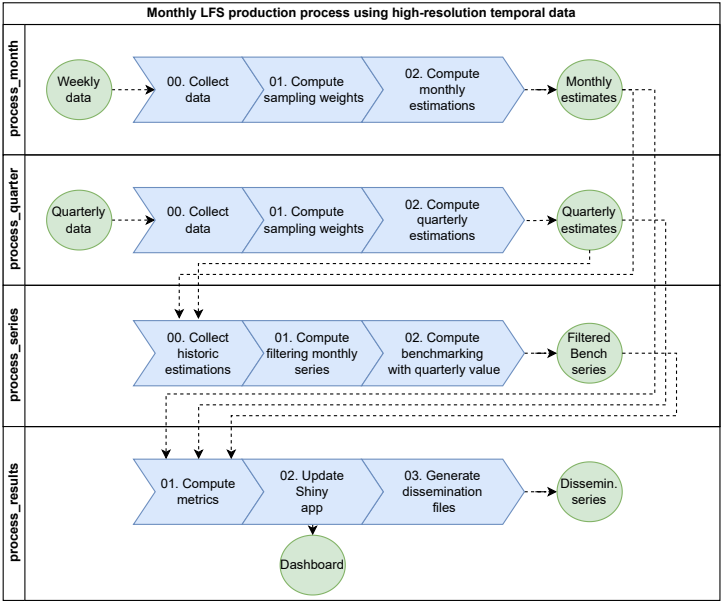
Use Case 1: Monthly LFS production process using high-resolution temporal data

- Novel end-to-end statistical production process that combines machine learning techniques, time series filtering, and benchmarking.
- Monthly (un)employment statistics:
  - Gender: male, female.
  - Age: 15-24, 25-74.
- LFS in Spain. Quarterly.
  - Weekly interviews.
  - Survey data: Quarterly sample.
- Significant improvement in time granularity.

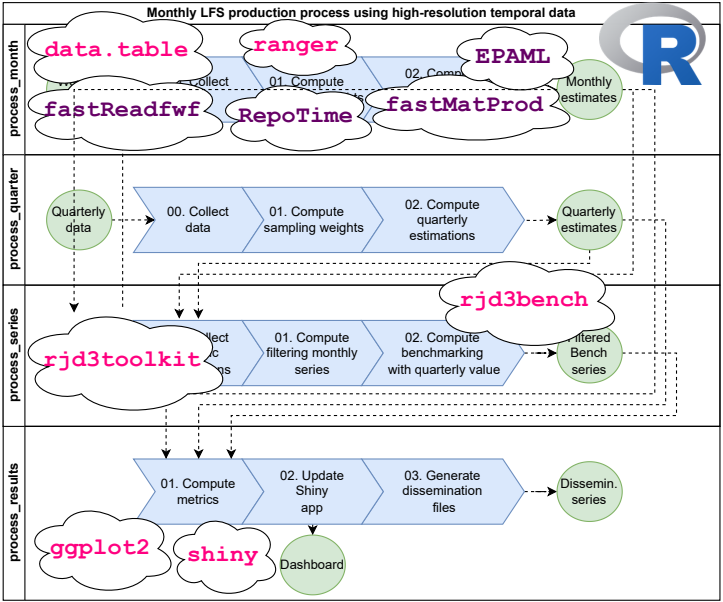
R as facilitator to implement an end-to-end statistical production process



R as facilitator to implement an end-to-end statistical production process



R as facilitator to implement an end-to-end statistical production process



High-resolution temporal data. Weekly Sampling Weights.

From quarterly to weekly.

1. Intermediate weights:

$$\omega_k^{\circ[W]} = \frac{N_{U_h}^{[M]}}{\sum_{k \in r_h^{[W]}} \frac{d_k^{[Q]}}{\pi_k^{[W \to Q]}} n_k} \frac{d_k^{[Q]}}{\pi_k^{[W \to Q]}}.$$

01. Compute sampling weights

where $\pi_k^{[W \to Q]}$ is the conditional inclusion probability for unit k to be interviewed in week W conditioned on its quarterly sample member.

$$\widehat{\pi}_k^{[W \to Q]} = \sum_{r=1}^{6} \mathbb{P}_k \left[(W, r) | S, P, R \right],$$

where r are the rotation group, S is the *stratum*, R is the *region* (NUTS2) and P is the *province* (NUTS3).

This conditional probability is estimated with a Random Forest model (see [Murphy, 2012]) using R package `ranger`.

2. Calibrated weights: $\omega_k^{[W]}$ computed using R package `fastMatProd` (in-house dev based on calib R package).

INE

High-resolution temporal data. Aggregates and variances.

- Weekly aggregate for week w: $\widehat{Y}_w^{[W]} = \sum_{k \in s_w} \omega_k^{[W]} y_k$.
- Monthly aggregate for month m based on weekly aggregates:

$$\widehat{Y}_m^{[M]} = \frac{1}{n_m^W} \sum_{w \in W_m} \widehat{Y}_w^{[W]}.$$

where W_m is the subset of weeks in month m, with cardinal n_m^W.

- Filtering time series with ARIMA-model-based decomposition: $Y = T + S_{13} + I$. ARIMA models for T, S_{13} and I are given by the canonical decomposition for the Y model. Filtering is done with the R package `rjd3toolkit`.
- Benchmarking is used once the quarterly estimates are computed to correct the provisional monthly estimates to be coherent with the corresponding quarterly estimate. Benchmarking is done with the R package `rjd3bench`.

INE

High-resolution temporal data. Aggregates and variances.

- Weekly aggregate for week w: $\widehat{Y}_w^{[W]} = \sum_{k \in s_w} \omega_k^{[W]} y_k$.
- Monthly aggregate for month m based on weekly aggregates:

$$\widehat{Y}_m^{[M]} = \frac{1}{n_m^W} \sum_{w \in W_m} \widehat{Y}_w^{[W]}.$$

02. Compute monthly estimations

where W_m is the subset of weeks in month m, with cardinal n_m^W.

- Filtering time series with A decomposition: $Y = T + S_{13} + I$. ARIMA models for T, S_{13} a e canonical decomposition for the Y model. Filtering is d ge `rjd3toolkit`.

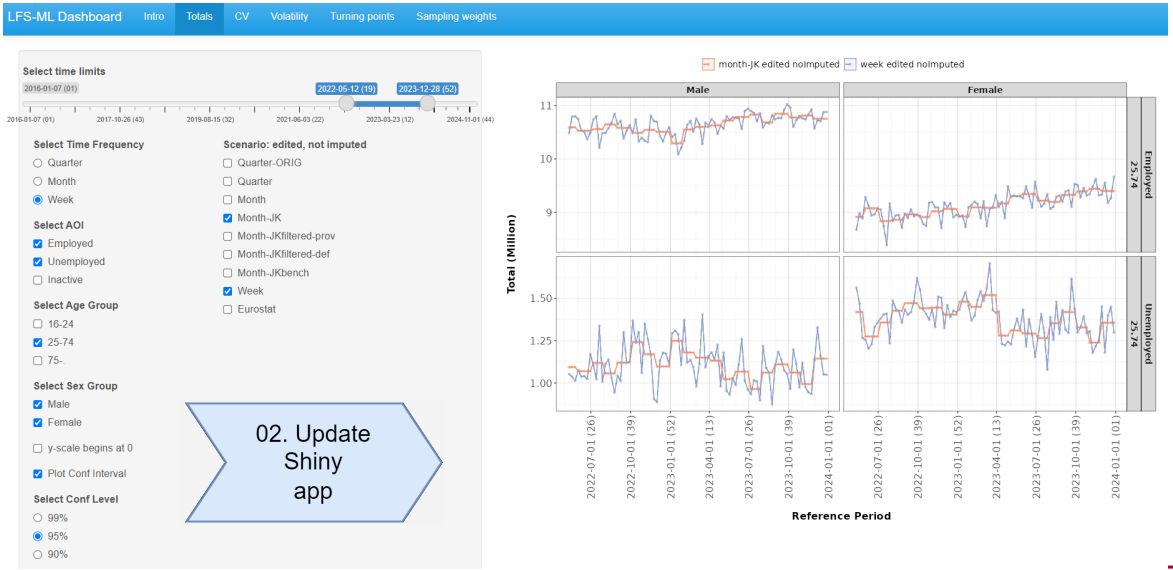
01. Compute filtering monthly series

- Benchmarking is used once the quarterly estimates are computed to correct the provisional monthly es nt with the corresponding quarterly estimate. Benchr the R package `rjd3bench`.

02. Compute benchmarking with quarterly value

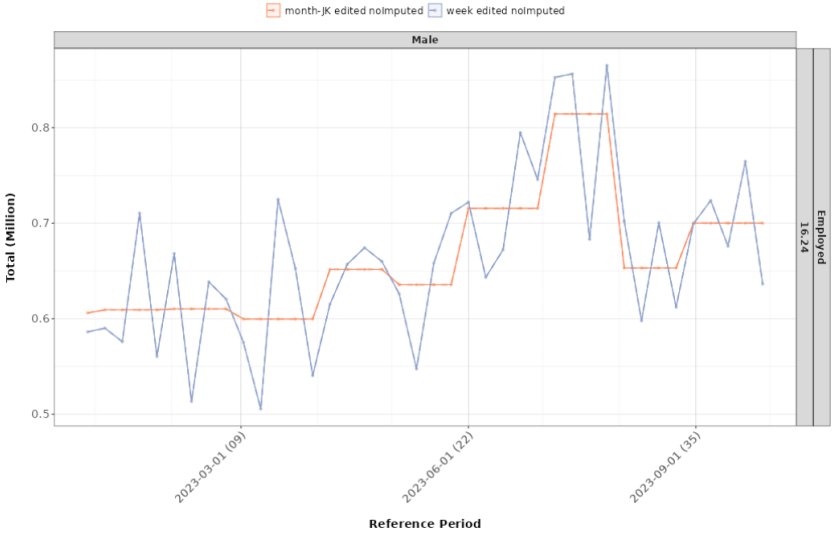
INE

R as facilitator to implement an end-to-end statistical production process

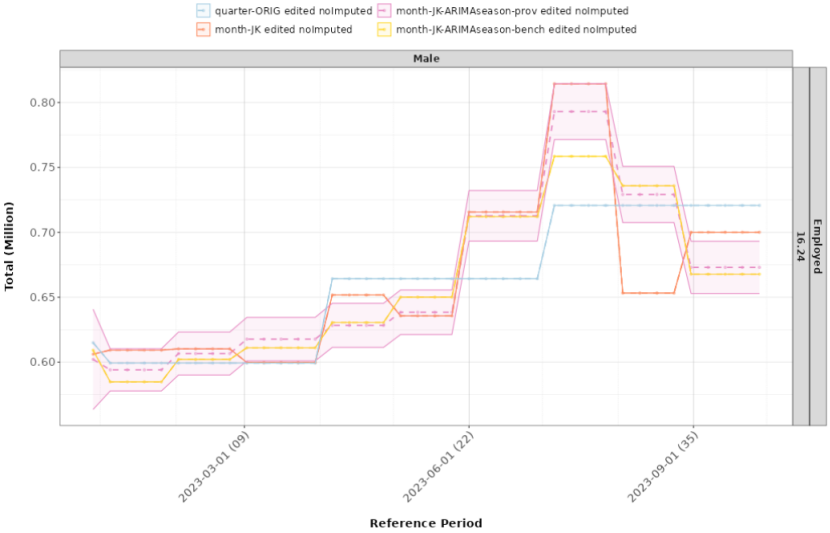


INE

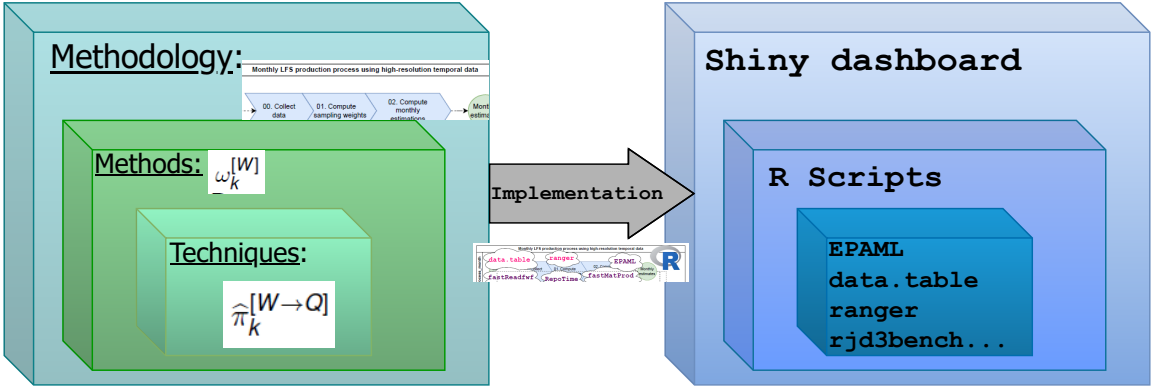
R as facilitator to implement an end-to-end statistical production process



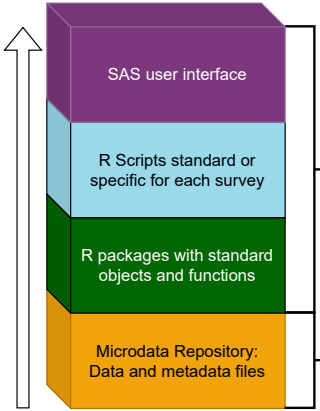
R as facilitator to implement an end-to-end statistical production process

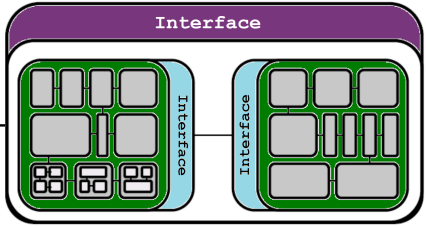


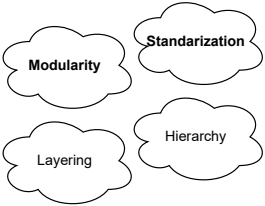
R as facilitator to implement an end-to-end statistical production process



Use Case 2: Ecosystem of R packages to interact with the Microdata Repository







Complete Key

Identifier: name of the statistical variable

Qualifiers:

In file name:

: (StatSpain internal) code of the statistical operation – E30103

: Validated value in *F*inal *F*ile in data dictionary *V*ersion *1* – FF_V1

: *M*onthly time period (*Jan 2016*) – MM012016

: *D*efinitive (complete) set of values *V*ersion *1* of the file/value – D_1

① : ID value of statistical unit – 289409300MM

② : NACE v2 code – 47.11

③ : (StatSpain internal) standardised NUTS2 code – 08

Value: value of the statistical variable – 9732

File name:

E30103.FF_V1.MM012016.D_1

Key-value pair:

$$\underbrace{\text{Turnover}}_{\text{Identifier}} + \underbrace{\overbrace{289409300\text{MM}}^{①}\overbrace{47.1}^{②}\overbrace{108}^{③}}_{\text{Qualifiers}} + \underbrace{9732}_{\text{Value}}$$

File-Internal Key

R as facilitator to connect with updated published data

Use Case 3: ineapir

R package `ineapir`: https://github.com/es-ine/ineapir

INE data extraction with ineapir : : CHEAT SHEET

The **ineapir** package allows to extract open data and metadata published by the INE (Spain). The data is obtained using calls to the INE API JSON service which access via URL requests to the data required by introducing the ID of the serie/tabla desired.

1 How to obtain ID's

Go to INE website and find a table/ series with the desired data (no need to be filtered). Depending the type of data hosted on the INE database we can distinguish several types of URL's:

FROM A TABLE

CASE 1: tempus type

The ID is the t parameter, **50902**

ine.es/jaxiT3/Tabla.htm?t=50902

CASE 2: pc- axis

The ID is the concatenation of path and file, **t20/e245/p08/l0/01001.px**

ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=01001.px

CASE 3: tpx

The ID is the tpx parameter, **33387**

https://www.ine.es/jaxi/Tabla.htm?tpx=33387&L=0

FROM A SERIES

CASE 4: series

1. Browse a table of interest
2. Filter the selected values
3. Click con the corresponding value cell
4. Take the code that appears on the plot

2 Main functions

OBTAINING DATA

- ***get_data_table(****idTable, filter, nlast, det, tip, lang, validate , verbose, unnest, metanames, metacodes)*
  It returns the data of the idTable specified according to the filter and the other arguments.
- ***get_data_series(****cdSeries,nlast, dateStart, dateEnd, det, tip, lang, validate , verbose, unnest)*
  It returns the data of the codSeries specified according to the parameters.
- ***get_data_series_filter(****operation, filter, periodicity, nlast, det,…)*
  It returns the data of the operation specified according to the filter and the other parameters.

Auxiliar functions for Tables

- ***get_metadata_table_groups/values(****idTable, (idGroup),…)*
  It returns all available groups and values for the specified table.
- ***get_metadata_table_ varval(****idTable,…)*
  Get metadata information about the variables and values for a given table
- ***get_metadata_series_table(****idTable, filter…)*
  Get all the series for a given table
- ***get_metadata_operation_table(****idTable,…)*
  It returns the operation for the specified table.

3 Arguments

OPTION	TYPE	DEFAULT	EFFECTS
idTable	int		id of the table

3a filter argument

Data from **tables** and **operations** can be filtered with a list according to the variables/values they contain. Let's see how to construct the filter:

FOR TABLES

See get_metadata_table_varval() to get all the values at once. There are different approaches to build the filter depending on the table type:

1. **tempus:** The filter is based on ids, with format: *list(id_variable1 = id_value1, id_variable2 = id_value2)*

Use Case 4: Common interface for Machine Learning algorithms

13^{00}– 14^{00} Scientific Session *Machine Learning and AI*
Chairman: Bogdan Oancea, University of Bucharest

1. *Supervised statistical (machine) learning for domain estimation with business survey data*
Vasilis Chasiotis, Department of Statistics, Athens University of Economics and Business, Athens, Greece

2. *The Synergy of R and Generative AI in Statistics*
Vytas Vaiciulis, Central Statistics Office, Ireland

3. *Automating Data Validation on SQL Server Using R and the Machine Learning package*
Paula Hartung, Statistics Iceland

4. Lightning talk: *RMLUtils: an Official Statistics oriented common interface for Machine Learning*
Luis Sanguiao Sande, Statistics Spain (INE)
Carlos Sáez Calvo, Statistics Spain (INE)
Sandra Barragán Andrés, Statistics Spain (INE)
Ester Puerto Sanz, Statistics Spain (INE)
María Novás Filgueira, Statistics Spain (INE)
Javier Villaescusa Almagro, Statistics Spain (INE)
Sergio Pardina Quirós, Statistics Spain (INE)
Álvaro García Tenorio, Statistics Spain (INE)

INE

Use Case 5: an R package for time series model identification

15^{30} - 16^{30} Scientific Session *NLP in Official Statistics and Time series*
Chairman: Mark Van Der Loo, Statistics Netherlands (CBS)

1. *A Neural Network Approach to Text Classification for International Standardized Codes*
Nina Niederhametner, Statistics Austria
Alexander Kowarik, Statistics Austria
Johannes Gussenbauer, Statistics Austria

2. *Semantic address matching using Keras for R*
Paula Cruz, Statistics Portugal; NOVA IMS
Leonardo Vanneschi, NOVA IMS
Marco Painho, NOVA IMS
Filipa Ribeiro, Statistics Portugal

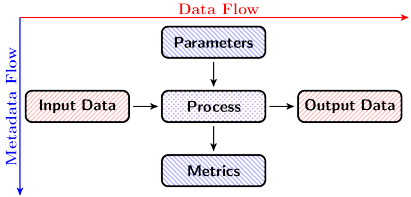
3. *TEAM: an R package for time series model identification*
C. Sáez Calvo, S.G. for Methodology and Sampling Design, Statistics Spain
L. Sanguiao Sande, S.G. for Methodology and Sampling Design, Statistics Spain
Félix Aparicio Pérez, S.G. for Methodology and Sampling Design, Statistics Spain
María Teresa Vázquez Gutiérrez, S.G. for Information Technologies and Communications, Statistics Spain
José Fernando Arranz Arauzo, S.G. for Information Technologies and Communications, Statistics Spain

Conclusions

INē

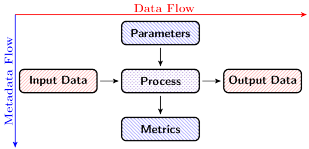
Conclusions: Take home messages

- R is open source, user friendly (easy to learn), constantly evolving. R has a bunch of packages implementing statistical methods and a wide community of users.
- R facilitates a proper **implementation** and **automation** of **modular processes** in the production of Official Statistics.

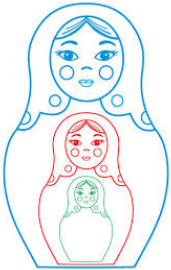


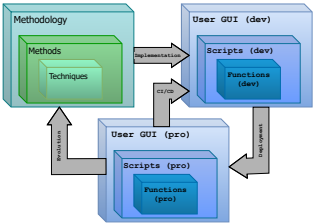
Conclusions: Take home messages

- Design the main process and subprocesses: **Modularity**



- Implement each subprocess: **Levels of implementation**





Run, or he's going to tell us about **R** again!


WHERE IS THAT R FREAK NOW?
imgflip.com

Thank you for your inspiration...

... and thank all of you here for your attention.

Questions?

sandra.barragan.andres@ine.es

References

Dustin Boswell and Trevor Foucher. *The Art of Readable Code: Simple and Practical Techniques for Writing Better Code*. " O'Reilly Media, Inc.", 2011.

Norman Matloff. *The art of R programming: A tour of statistical software design*. No Starch Press, 2011.

K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, New York, 2012.

D. Salgado, M E. Esteban, M. Novás, S. Saldaña, and L. Sanguiao. Data organisation and process design based on functional modularity for a standard production process. *Journal of official statistics*, 34(4):811–833, 2018.

Allen B Tucker and Robert Noonan. *Programming languages: principles and paradigms*. McGraw-Hill, 2007.

Mark PJ van der Loo. A method for deriving information from running r code. *arXiv preprint arXiv:2002.07472*, 2020.

INE