Implementing open science

Jonas Schöley

@jschoeley

0000-0002-3340-8518

schoeley@demogr.mpg.de

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Benefits of reproducibility and open science

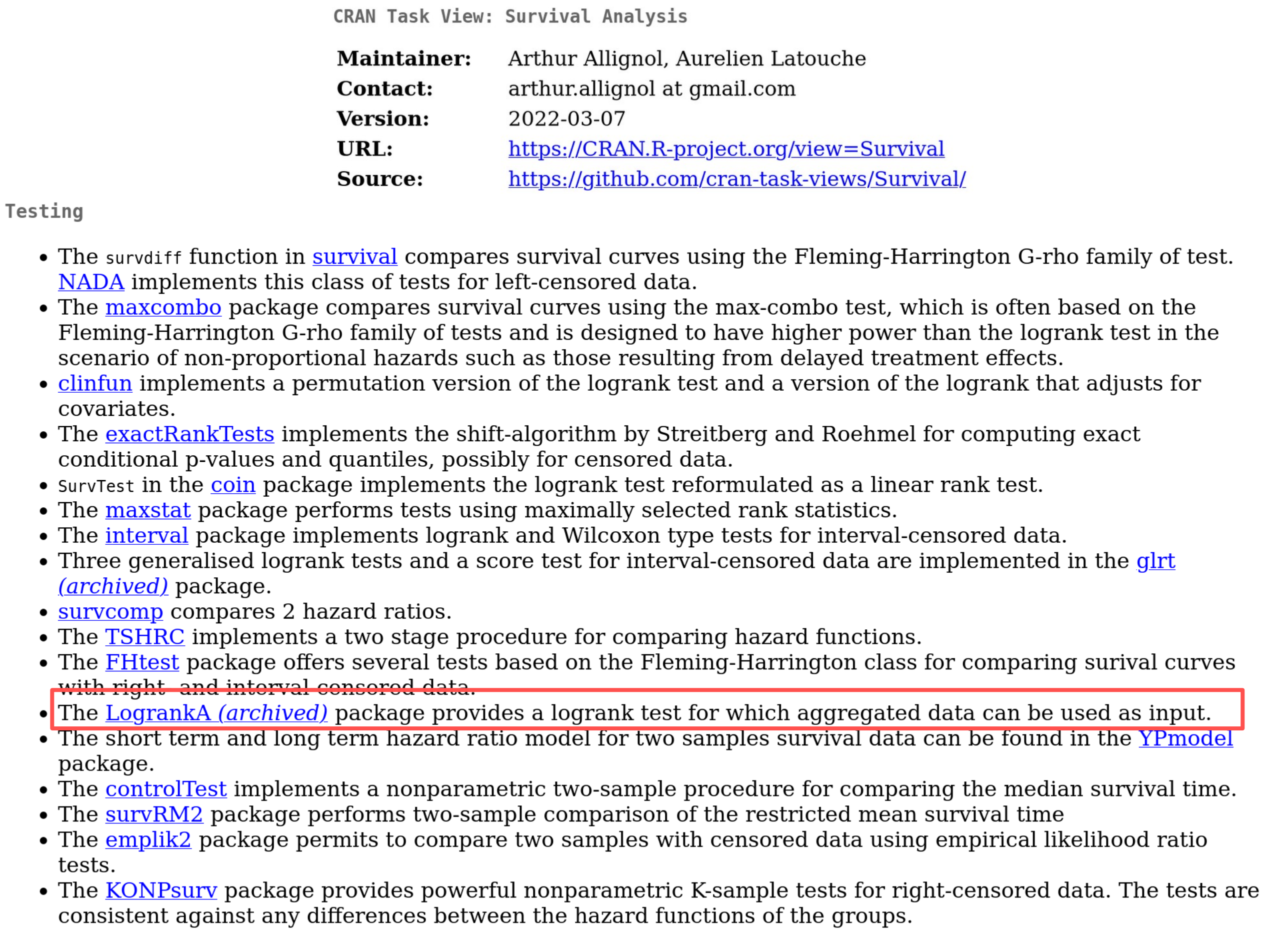
Some personal experiences

LogrankA

Together with

Roland **Rau**

@Demographie



CRAN Task View: Survival Analysis

Maintainer: Arthur Allignol, Aurelien Latouche
Contact: arthur.allignol at gmail.com
Version: 2022-03-07
URL: https://CRAN.R-project.org/view=Survival
Source: https://github.com/cran-task-views/Survival/

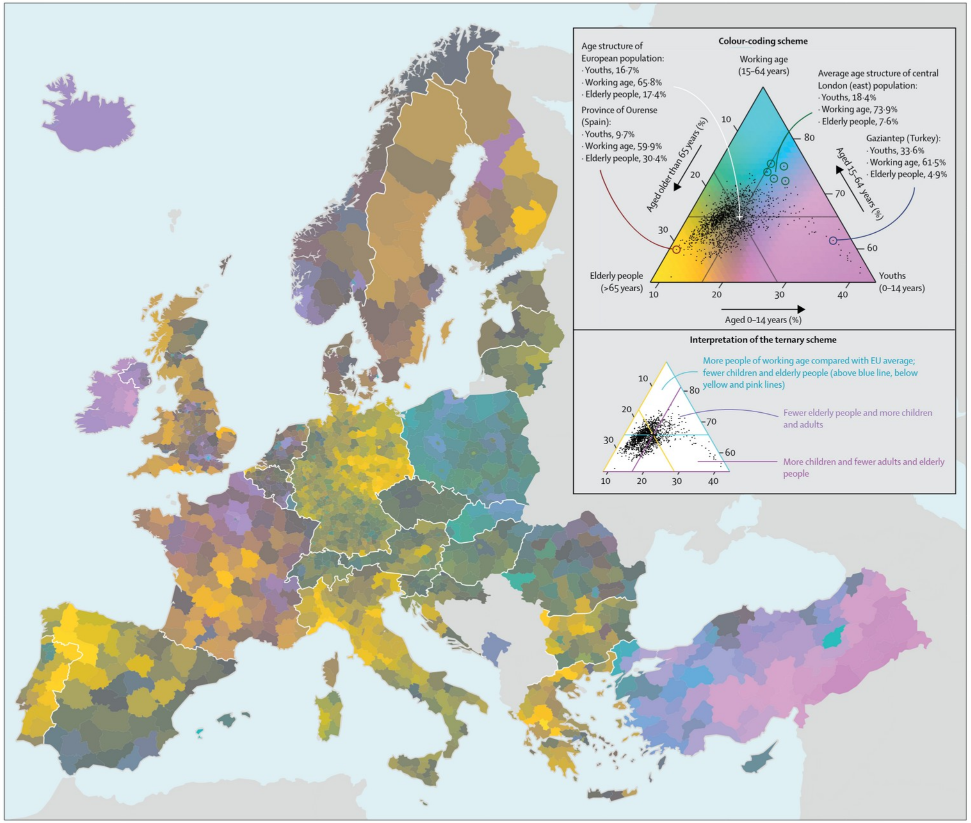
Testing

- The `survdiff` function in survival compares survival curves using the Fleming-Harrington G-rho family of test. NADA implements this class of tests for left-censored data.
- The maxcombo package compares survival curves using the max-combo test, which is often based on the Fleming-Harrington G-rho family of tests and is designed to have higher power than the logrank test in the scenario of non-proportional hazards such as those resulting from delayed treatment effects.
- clinfun implements a permutation version of the logrank test and a version of the logrank that adjusts for covariates.
- The exactRankTests implements the shift-algorithm by Streitberg and Roehmel for computing exact conditional p-values and quantiles, possibly for censored data.
- `SurvTest` in the coin package implements the logrank test reformulated as a linear rank test.
- The maxstat package performs tests using maximally selected rank statistics.
- The interval package implements logrank and Wilcoxon type tests for interval-censored data.
- Three generalised logrank tests and a score test for interval-censored data are implemented in the glrt *(archived)* package.
- survcomp compares 2 hazard ratios.
- The TSHRC implements a two stage procedure for comparing hazard functions.
- The FHtest package offers several tests based on the Fleming-Harrington class for comparing surival curves with right- and interval-censored data.
- The LogrankA *(archived)* package provides a logrank test for which aggregated data can be used as input.
- The short term and long term hazard ratio model for two samples survival data can be found in the YPmodel package.
- The controlTest implements a nonparametric two-sample procedure for comparing the median survival time.
- The survRM2 package performs two-sample comparison of the restricted mean survival time
- The emplik2 package permits to compare two samples with censored data using empirical likelihood ratio tests.
- The KONPsurv package provides powerful nonparametric K-sample tests for right-censored data. The tests are consistent against any differences between the hazard functions of the groups.

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky



Colour-coding scheme
Age structure of European population:
· Youths, 16·7%
· Working age, 65·8%
· Elderly people, 17·4%
Province of Ourense (Spain):
· Youths, 9·7%
· Working age, 59·9%
· Elderly people, 30·4%
Working age (15–64 years)
Average age structure of central London (east) population:
· Youths, 18·4%
· Working age, 73·9%
· Elderly people, 7·6%
Gaziantep (Turkey):
· Youths, 33·6%
· Working age, 61·5%
· Elderly people, 4·9%
Aged older than 65 years (%)
Aged 15–64 years (%)
Aged 0–14 years (%)
Elderly people (>65 years)
Youths (0–14 years)
10
20
30
40
60
70
80
Interpretation of the ternary scheme
More people of working age compared with EU average; fewer children and elderly people (above blue line, below yellow and pink lines)
Fewer elderly people and more children and adults
More children and fewer adults and elderly people

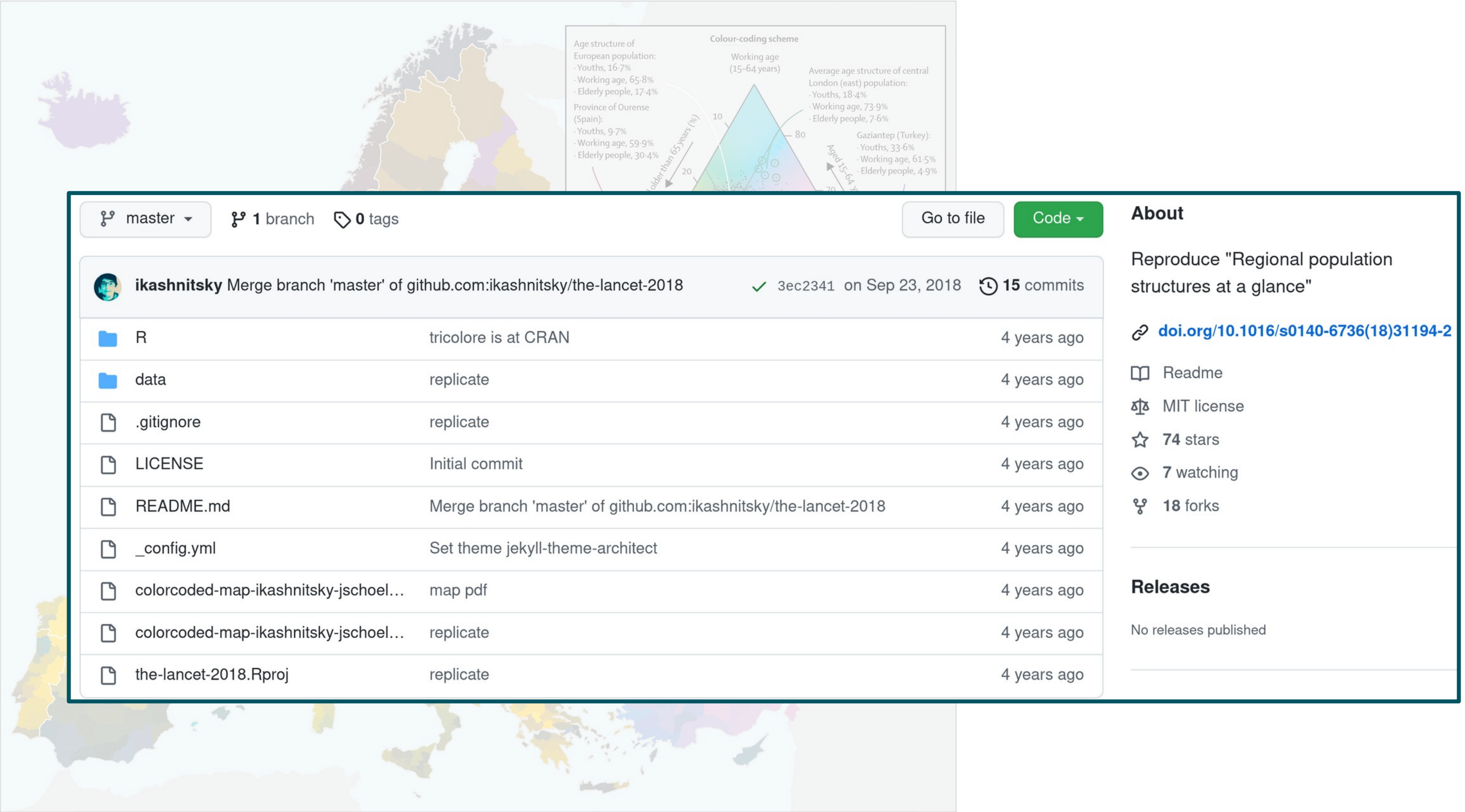
***Figure:* Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 2015**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of deviations from the centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encodes the direction of deviation: yellow indicates an elderly population (>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightness components signify the distance from the centre ranging from desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.[10]

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky



ikashnitsky Merge branch 'master' of github.com:ikashnitsky/the-lancet-2018 ✓ 3ec2341 on Sep 23, 2018 15 commits

R	tricolore is at CRAN	4 years ago
data	replicate	4 years ago
.gitignore	replicate	4 years ago
LICENSE	Initial commit	4 years ago
README.md	Merge branch 'master' of github.com:ikashnitsky/the-lancet-2018	4 years ago
_config.yml	Set theme jekyll-theme-architect	4 years ago
colorcoded-map-ikashnitsky-jschoel…	map pdf	4 years ago
colorcoded-map-ikashnitsky-jschoel…	replicate	4 years ago
the-lancet-2018.Rproj	replicate	4 years ago

About

Reproduce "Regional population structures at a glance"

doi.org/10.1016/s0140-6736(18)31194-2

Readme

MIT license

74 stars

7 watching

18 forks

Releases

No releases published

Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 2015**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of deviations from the centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encodes the direction of deviation: yellow indicates an elderly population (>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightness components signify the distance from the centre ranging from desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.[10]

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky

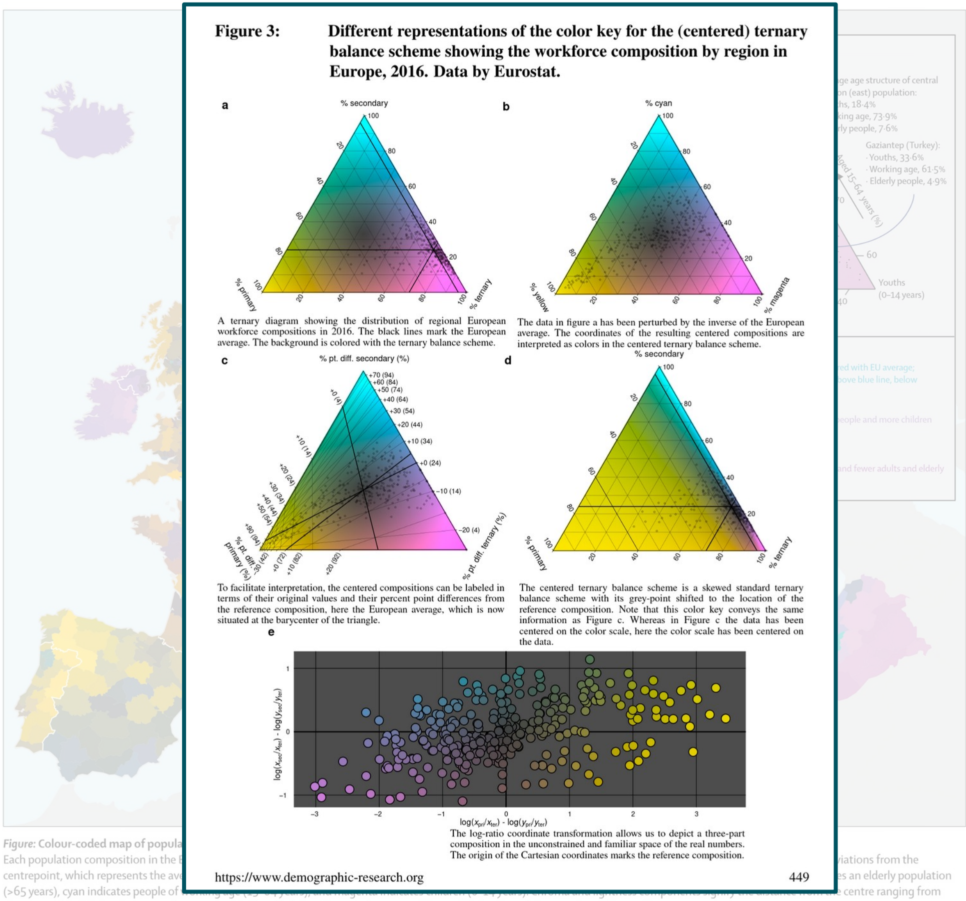


Figure 3: Different representations of the color key for the (centered) ternary balance scheme showing the workforce composition by region in Europe, 2016. Data by Eurostat.

A ternary diagram showing the distribution of regional European workforce compositions in 2016. The black lines mark the European average. The background is colored with the ternary balance scheme.

The data in figure a has been perturbed by the inverse of the European average. The coordinates of the resulting centered compositions are interpreted as colors in the centered ternary balance scheme.

To facilitate interpretation, the centered compositions can be labeled in terms of their original values and their percent point differences from the reference composition, here the European average, which is now situated at the barycenter of the triangle.

The centered ternary balance scheme is a skewed standard ternary balance scheme with its grey-point shifted to the location of the reference composition. Note that this color key conveys the same information as Figure c. Whereas in Figure c the data has been centered on the color scale, here the color scale has been centered on the data.

e
$\log(x_{sec}/x_{ter}) - \log(y_{sec}/y_{ter})$
$\log(x_{pri}/x_{ter}) - \log(y_{pri}/y_{ter})$

The log-ratio coordinate transformation allows us to depict a three-part composition in the unconstrained and familiar space of the real numbers. The origin of the Cartesian coordinates marks the reference composition.

https://www.demographic-research.org 449

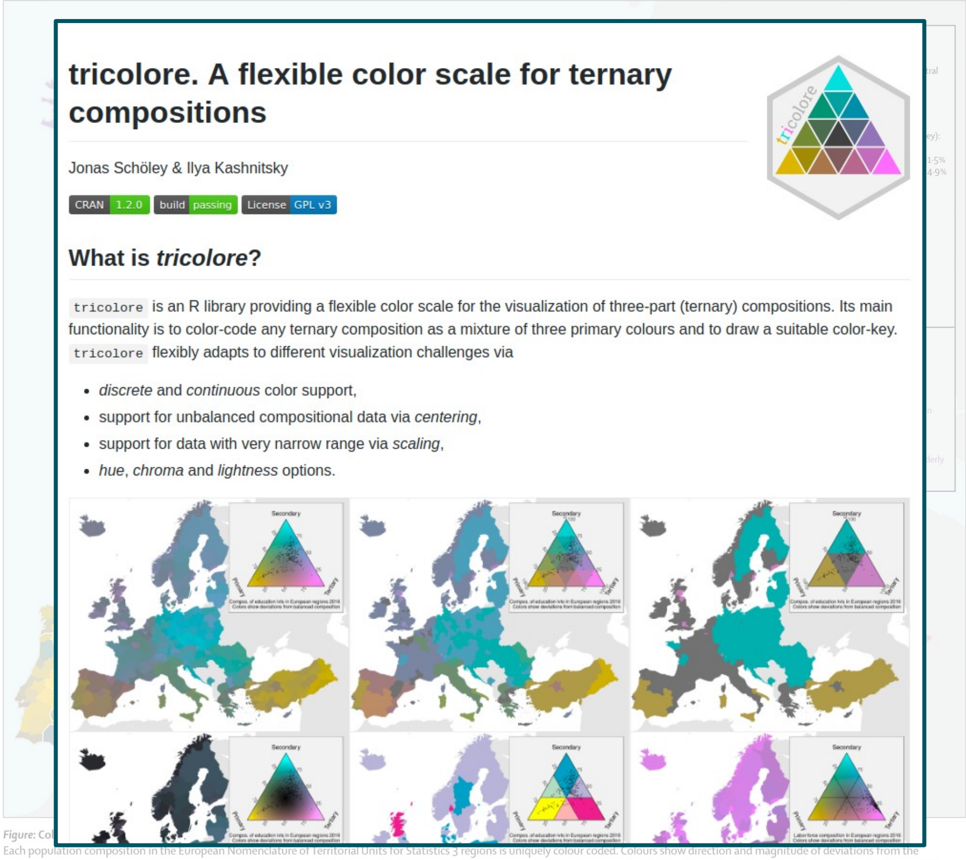
Figure: Colour-coded map of popula
Each population composition in the E … viations from the
centrepoint, which represents the ave … es an elderly population
(>65 years), cyan indicates people of … e centre ranging from
desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.[10]

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky



tricolore. A flexible color scale for ternary compositions

Jonas Schöley & Ilya Kashnitsky

CRAN 1.2.0 build passing License GPL v3

What is *tricolore*?

`tricolore` is an R library providing a flexible color scale for the visualization of three-part (ternary) compositions. Its main functionality is to color-code any ternary composition as a mixture of three primary colours and to draw a suitable color-key. `tricolore` flexibly adapts to different visualization challenges via

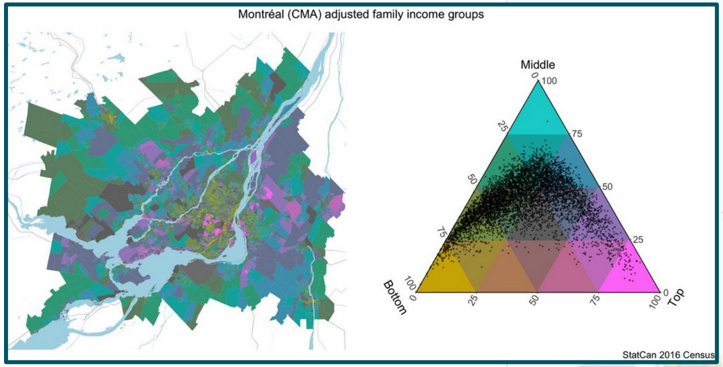
- *discrete* and *continuous* color support,
- support for unbalanced compositional data via *centering*,
- support for data with very narrow range via *scaling*,
- *hue*, *chroma* and *lightness* options.

Figure: Col
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of deviations from the centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encodes the direction of deviation: yellow indicates an elderly population (>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightness components signify the distance from the centre ranging from desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.[10]

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

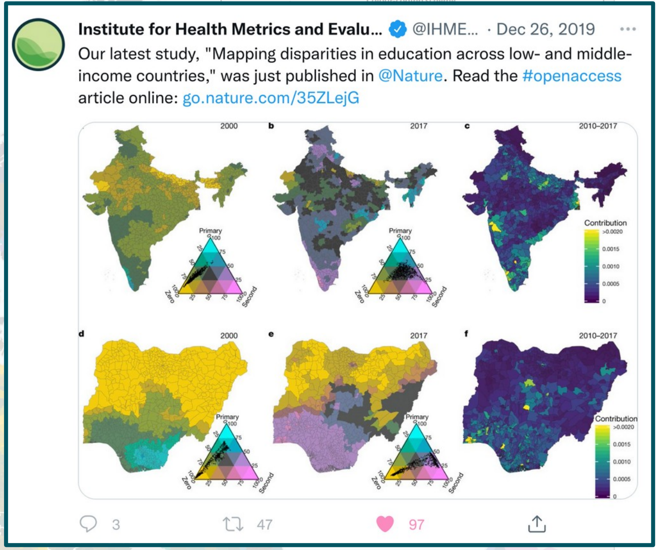
tricolore

Income distribution in Canadian cities



Montréal (CMA) adjusted family income groups

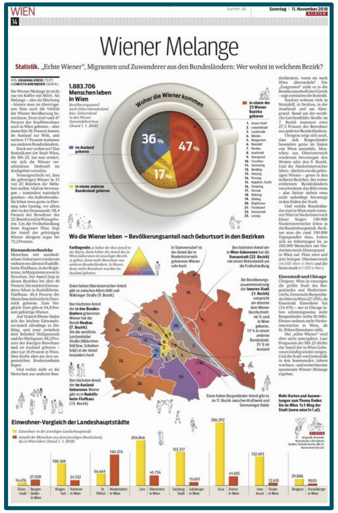
LMIC education disparity



Institute for Health Metrics and Evalu... @IHME... · Dec 26, 2019

Our latest study, "Mapping disparities in education across low- and middle-income countries," was just published in @Nature. Read the #openaccess article online: go.nature.com/35ZLejG

Vienna's population by origin



WIEN 14 kurier.at Sonntag | 11. November 2018 KURIER

Wiener Melange

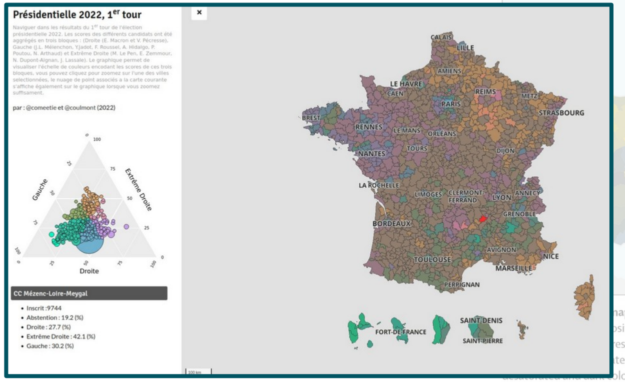
Statistik. „Echte Wiener", Migranten und Zuwanderer aus den Bundesländern: Wer wohnt in welchem Bezirk?

VON JOHANNA KREID (TEXT) & CHRISTA BREINEDER (GRAFIK)

Die Wiener Melange ist nicht nur ein Kaffee mit Milch. Als Melange – also als Mischung – könnte man im übertragenen Sinn auch die Vielfalt der Wiener Bevölkerung bezeichnen: Zwar sind rund 47 Prozent der Stadtbewohner auch in Wien geboren – aber immerhin 36 Prozent kamen im Ausland zur Welt, und weitere 17 Prozent stammen

1.883.706 Menschen leben in Wien
Woher die Wiener kommen
36 %
47 %
17 %
in einem der 23 Wiener Bezirke geboren
im Ausland geboren
in einem anderen Bundesland geboren
Wo die Wiener leben – Bevölkerungsanteil nach Geburtsort in den Bezirken
Einwohner-Vergleich der Landeshauptstädte

French election results



Présidentielle 2022, 1^er tour

Naviguer dans les résultats du 1^er tour de l'élection présidentielle 2022. Les scores des différents candidats ont été aggrégés en trois bloques : (Droite (E. Macron et V. Pécresse), Gauche (J.L. Mélenchon, Y.Jadot, F. Roussel, A. Hidalgo, P. Poutou, N. Arthaud) et Extrême Droite (M. Le Pen, E. Zemmour, N. Dupont-Aignan, J. Lassale). Le graphique permet de visualiser l'échelle de couleurs encodant les scores de ces trois bloques, vous pouvez cliquez pour zoomez sur l'une des villes selectionnées, le nuage de point associés a la carte courante s'affiche également sur le graphique lorsque vous zoomez suffisament.

par : @comeetie et @coulmont (2022)

CC Mézenc-Loire-Meygal

- Inscrit :9744
- Abstention : 19.2 (%)
- Droite : 27.7 (%)
- Extrême Droite : 42.1 (%)
- Gauche : 30.2 (%)

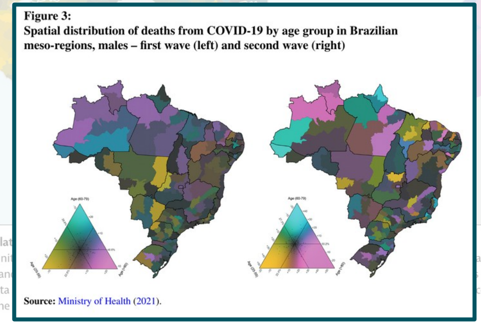


Figure 3:
Spatial distribution of deaths from COVID-19 by age group in Brazilian meso-regions, males – first wave (left) and second wave (right)

Source: Ministry of Health (2021).

Regional age distribution of COVID deaths In Brazil

nap of population structures in European Nomenclat
osition in the European Nomenclature of Territorial Unit
resents the average age of the European population, an
ates people of working age (15–64 years), and magenta
olours near the centre to vivid and bright colours at the
ations from the
s an elderly population
centre ranging from

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Agricultural and Forest Meteorology

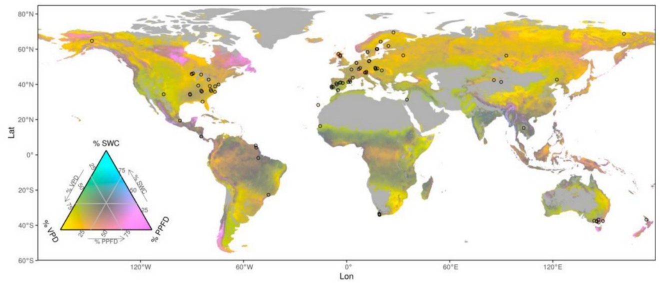


Fig. 3. Relative importance (relative R^2) of the three hydrometeorological drivers of transpiration regulation. Relative R^2 were calculated at each cell dividing the projection of each three drivers partial R^2 from the FULL model by the sum of the three partial R^2 at the same cell. Partial R^2 were projected at the global scale using linear models with climate, soil and vegetation structural variables as explanatory variables (Table S6). Grid colours were calculated using the 'tricolore' package (Schöley and Kashnitsky, 2020) for each cell. Colour gradient indicate the relative importance of the three hydrometeorological constraints. Light grey colour are non-forested areas or with vegetation not taller than 0.5 m (Simard et al., 2011). % VPD: vapour pressure deficit relative importance. % SWC: soil water content relative importance. % PPFD: photosynthetic photon flux density relative importance. Points indicate locations of study sites.

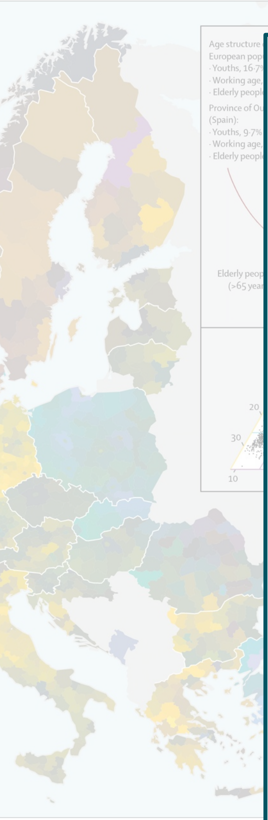


Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 20**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Co
centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encode
(>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightness
desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.

Soil composition

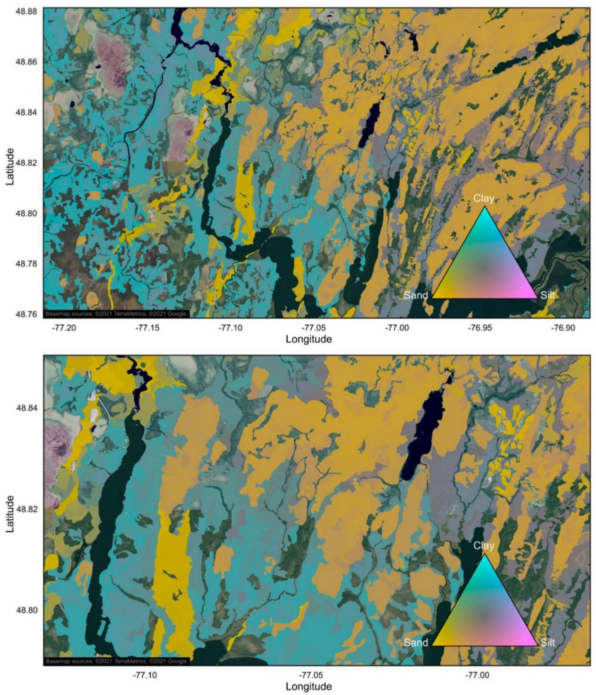


Figure 8 **Maps of soil texture composition at two zoom levels, for a region at the southern edge of the Abitibi and James Bay Lowlands soil province.** Only productive forest land characterized by mineral soils was mapped. Agricultural and unproductive forest land, organic soils, anthropogenic infrastructures, and water areas were excluded. Maps were produced with QGIS software, version 3.4 (*QGIS, 2020*). Basemap credit: ©2021 TerraMetrics, ©2021 Google.

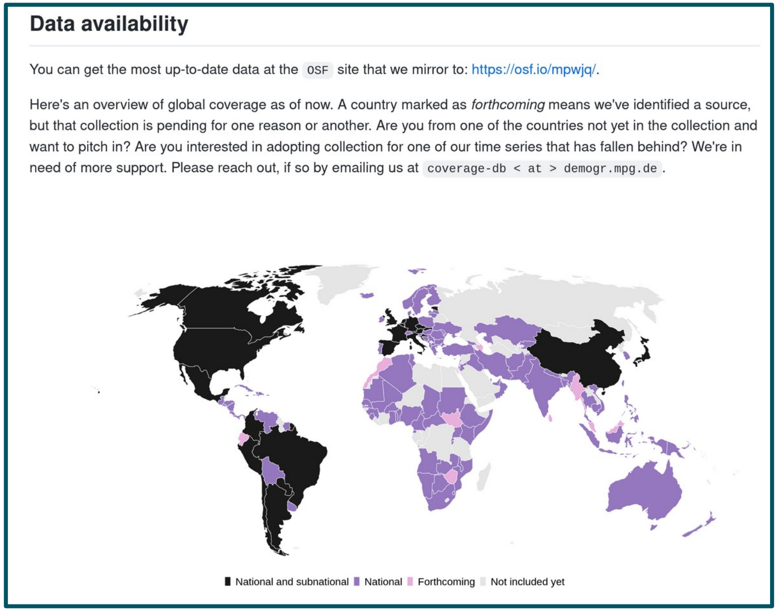
Full-size DOI: 10.7717/peerj.11685/fig-8

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

Reproducible Demographers

Tim **Riffe** @timriffe1 & Enrique **Acosta** @Acosta_Kike_ &
Manal **Elzalabany** & Maxi **Kniffka** @MaxiKniffka & Jessica **Donzowa** @jdonzowa

Created a fully reproducible data base of age specific COVID-19 statistics.



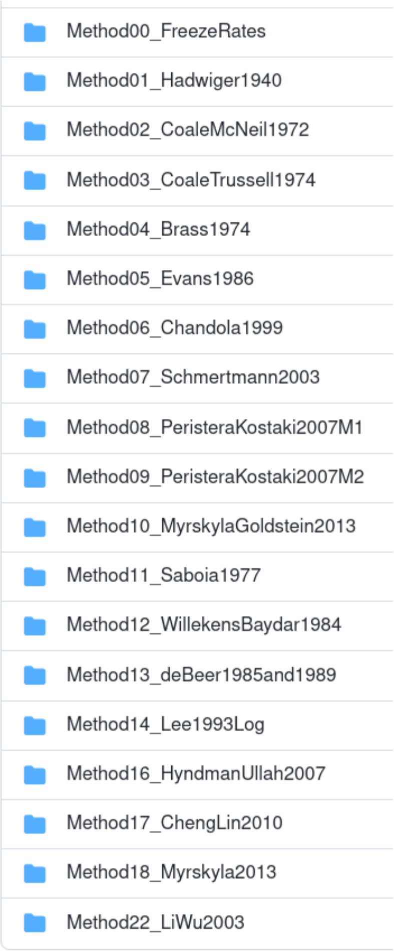
Data availability

You can get the most up-to-date data at the `OSF` site that we mirror to: https://osf.io/mpwjq/.

Here's an overview of global coverage as of now. A country marked as *forthcoming* means we've identified a source, but that collection is pending for one reason or another. Are you from one of the countries not yet in the collection and want to pitch in? Are you interested in adopting collection for one of our time series that has fallen behind? We're in need of more support. Please reach out, if so by emailing us at `coverage-db < at > demogr.mpg.de`.

Reproducible Demographers

Christina **Bohk**



- Method00_FreezeRates
- Method01_Hadwiger1940
- Method02_CoaleMcNeil1972
- Method03_CoaleTrussell1974
- Method04_Brass1974
- Method05_Evans1986
- Method06_Chandola1999
- Method07_Schmertmann2003
- Method08_PeristeraKostaki2007M1
- Method09_PeristeraKostaki2007M2
- Method10_MyrskylaGoldstein2013
- Method11_Saboia1977
- Method12_WillekensBaydar1984
- Method13_deBeer1985and1989
- Method14_Lee1993Log
- Method16_HyndmanUllah2007
- Method17_ChengLin2010
- Method18_Myrskyla2013
- Method22_LiWu2003

Bohk et al. (2018). Forecast accuracy hardly improves with methods complexity when completing cohort fertility. 10.1109/5.771073

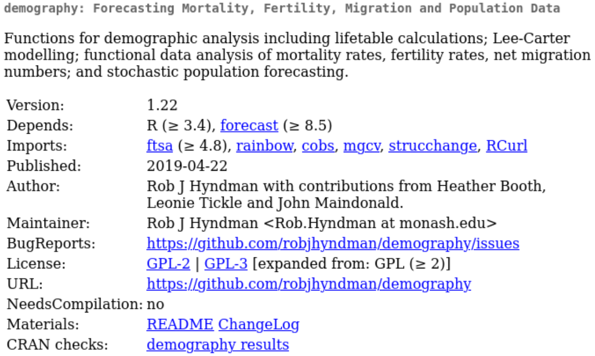
Implemented, shared and compared 22 fertility forecasting methods

github.com/fertility-forecasting/validate-forecast-methods/tree/master/basic-scripts-forecast-methods

Reproducible Demographers

Rob **Hyndman**

Author of countless R packages widely applied in demography.



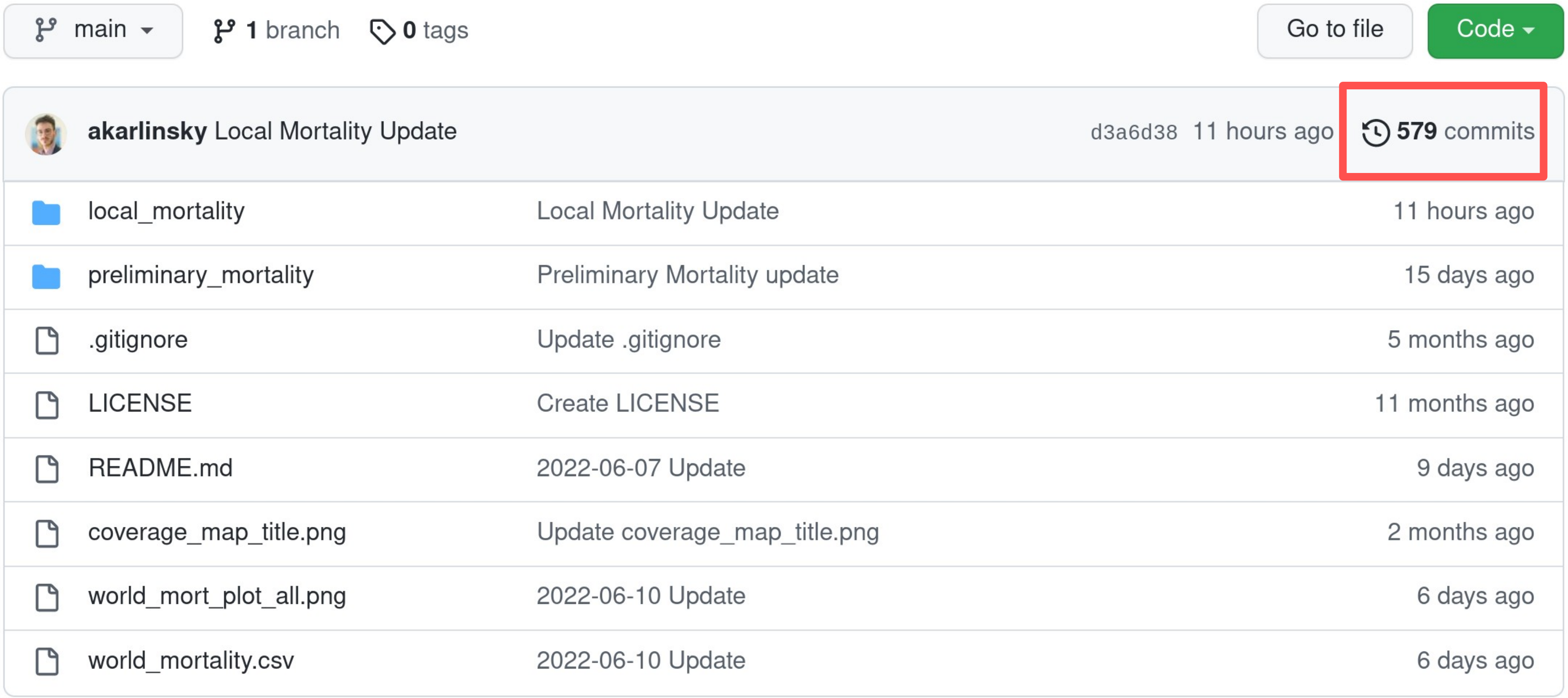
demography: Forecasting Mortality, Fertility, Migration and Population Data

Functions for demographic analysis including lifetable calculations; Lee-Carter modelling; functional data analysis of mortality rates, fertility rates, net migration numbers; and stochastic population forecasting.

Version:	1.22
Depends:	R (≥ 3.4), forecast (≥ 8.5)
Imports:	ftsa (≥ 4.8), rainbow, cobs, mgcv, strucchange, RCurl
Published:	2019-04-22
Author:	Rob J Hyndman with contributions from Heather Booth, Leonie Tickle and John Maindonald.
Maintainer:	Rob J Hyndman <Rob.Hyndman at monash.edu>
BugReports:	https://github.com/robjhyndman/demography/issues
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL:	https://github.com/robjhyndman/demography
NeedsCompilation:	no
Materials:	README ChangeLog
CRAN checks:	demography results

github.com/robjhyndman

Consuming open science

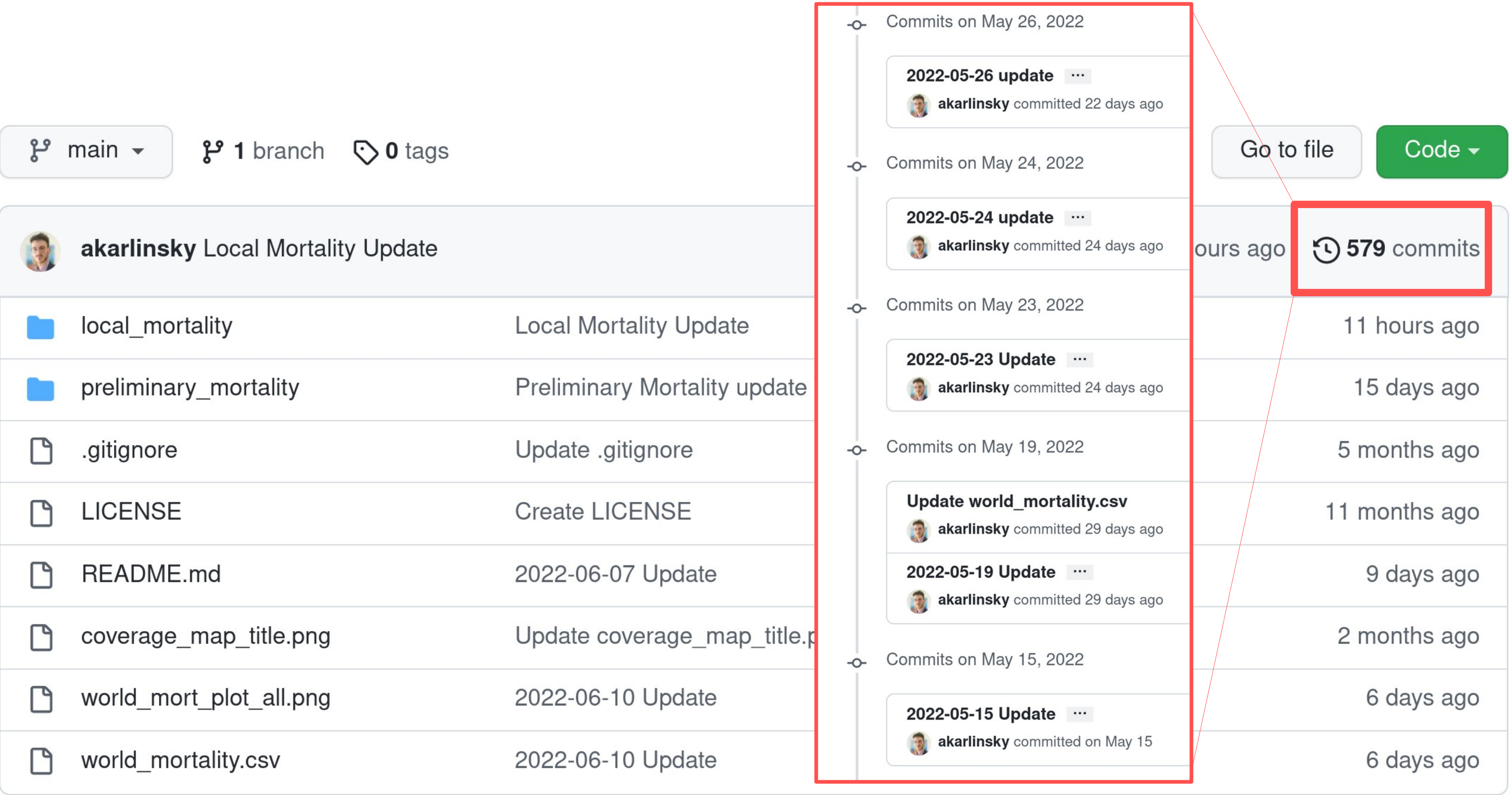


akarlinsky Local Mortality Update d3a6d38 11 hours ago **579** commits

local_mortality	Local Mortality Update	11 hours ago
preliminary_mortality	Preliminary Mortality update	15 days ago
.gitignore	Update .gitignore	5 months ago
LICENSE	Create LICENSE	11 months ago
README.md	2022-06-07 Update	9 days ago
coverage_map_title.png	Update coverage_map_title.png	2 months ago
world_mort_plot_all.png	2022-06-10 Update	6 days ago
world_mortality.csv	2022-06-10 Update	6 days ago

Karlinsky & Kobak (2022). World Mortality Database. github.com/akarlinsky/world_mortality

Consuming open science



main · 1 branch · 0 tags · Go to file · Code

akarlinsky Local Mortality Update · ours ago · 579 commits

local_mortality	Local Mortality Update	11 hours ago
preliminary_mortality	Preliminary Mortality update	15 days ago
.gitignore	Update .gitignore	5 months ago
LICENSE	Create LICENSE	11 months ago
README.md	2022-06-07 Update	9 days ago
coverage_map_title.png	Update coverage_map_title.p	2 months ago
world_mort_plot_all.png	2022-06-10 Update	6 days ago
world_mortality.csv	2022-06-10 Update	6 days ago

Commits on May 26, 2022

2022-05-26 update
akarlinsky committed 22 days ago

Commits on May 24, 2022

2022-05-24 update
akarlinsky committed 24 days ago

Commits on May 23, 2022

2022-05-23 Update
akarlinsky committed 24 days ago

Commits on May 19, 2022

Update world_mortality.csv
akarlinsky committed 29 days ago

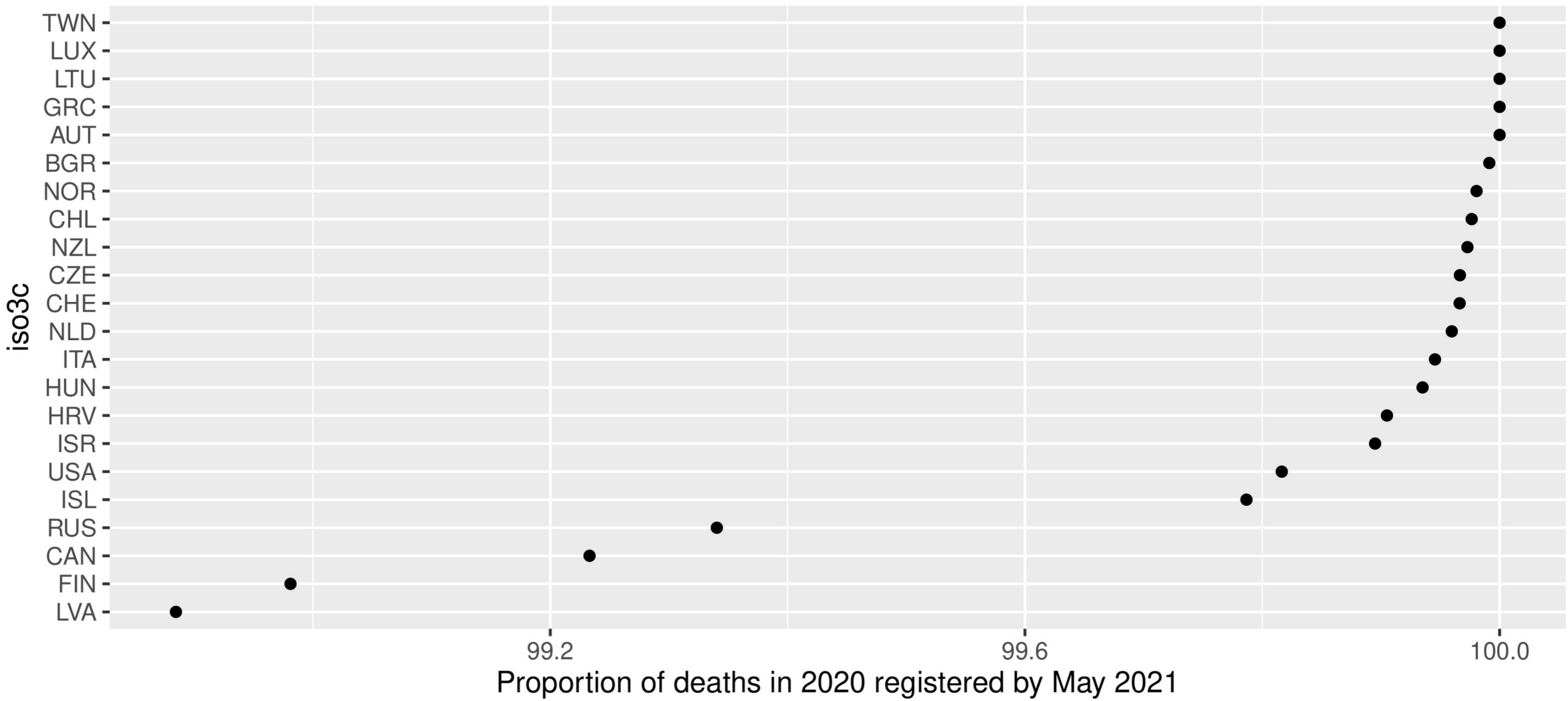
2022-05-19 Update
akarlinsky committed 29 days ago

Commits on May 15, 2022

2022-05-15 Update
akarlinsky committed on May 15

Karlinsky & Kobak (2022). World Mortality Database. github.com/akarlinsky/world_mortality

Consuming open science



Derived from Karlinsky & Kobak (2022). World Mortality Database. github.com/akarlinsky/world_mortality

Reproducibility & replicability

Given **your data** and **your analysis**
I arrive at **your results**

Given **your research question**,
my data and **my analysis**
I arrive at **your results**

Implementing reproducibility & replicability

Ensure everyone can run your analysis

R

Share and version your analysis

git

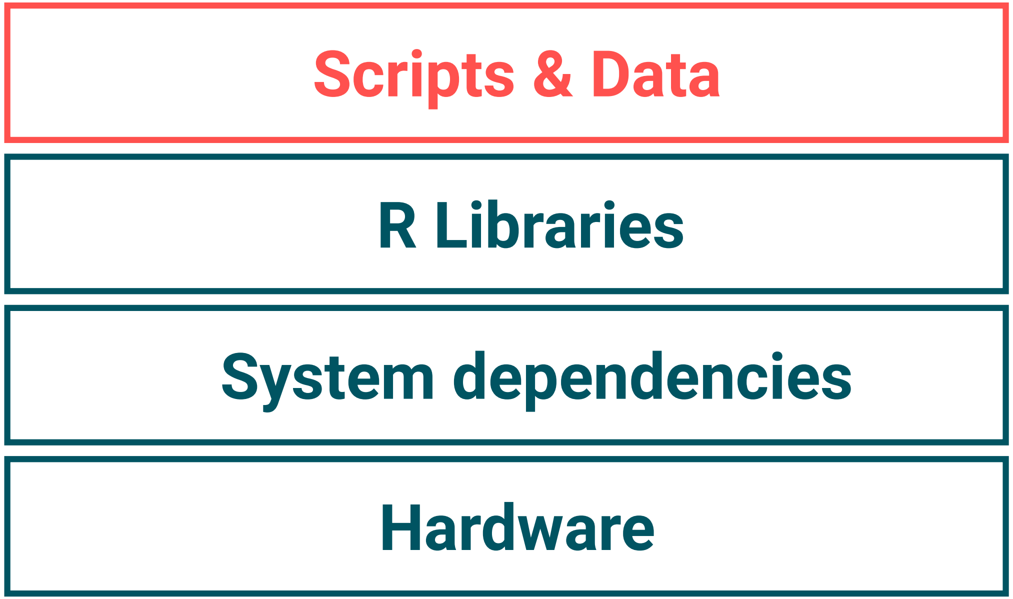
GitHub

Share your data Archive your data Get DOIs

zenodo

The computational reproducibility stack

Ensure everyone can run your analysis



Scripts & Data

R Libraries

System dependencies

Hardware

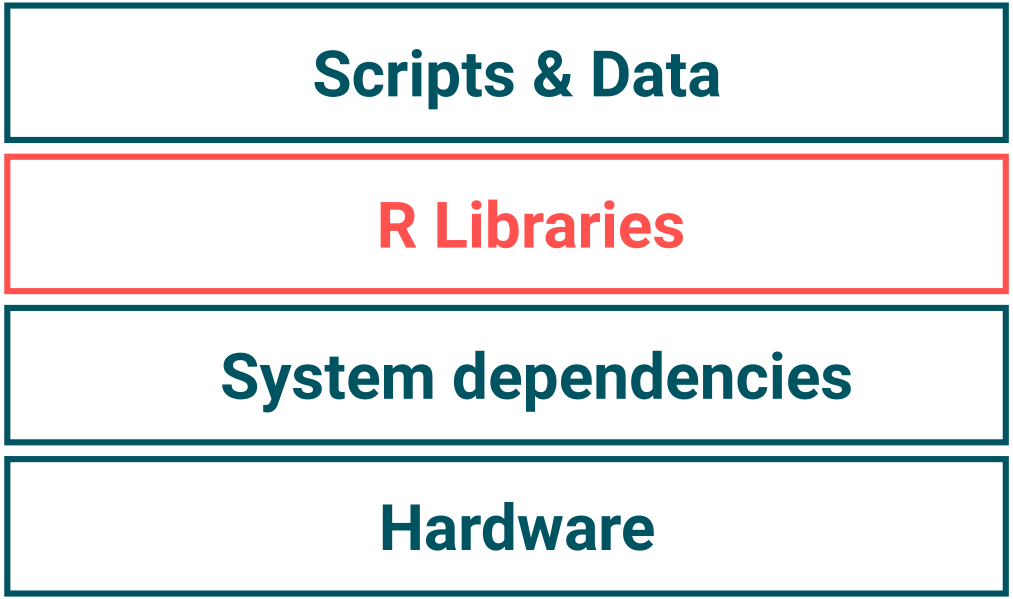
git zenodo

renv

docker

The computational reproducibility stack

Ensure everyone can run your analysis



Scripts & Data

R Libraries

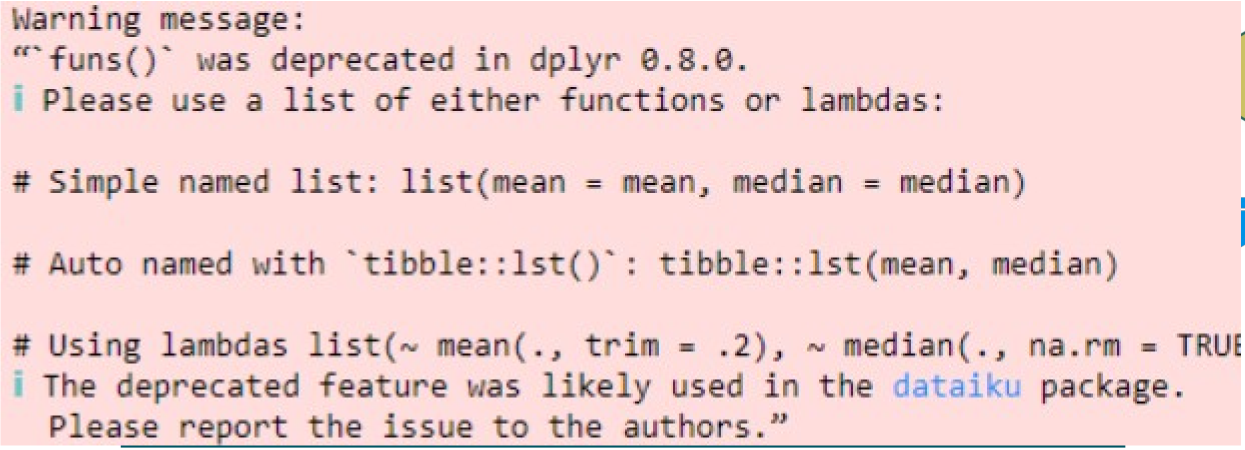
System dependencies

Hardware

The computational reproducibility stack

Ensure everyone can run your analysis

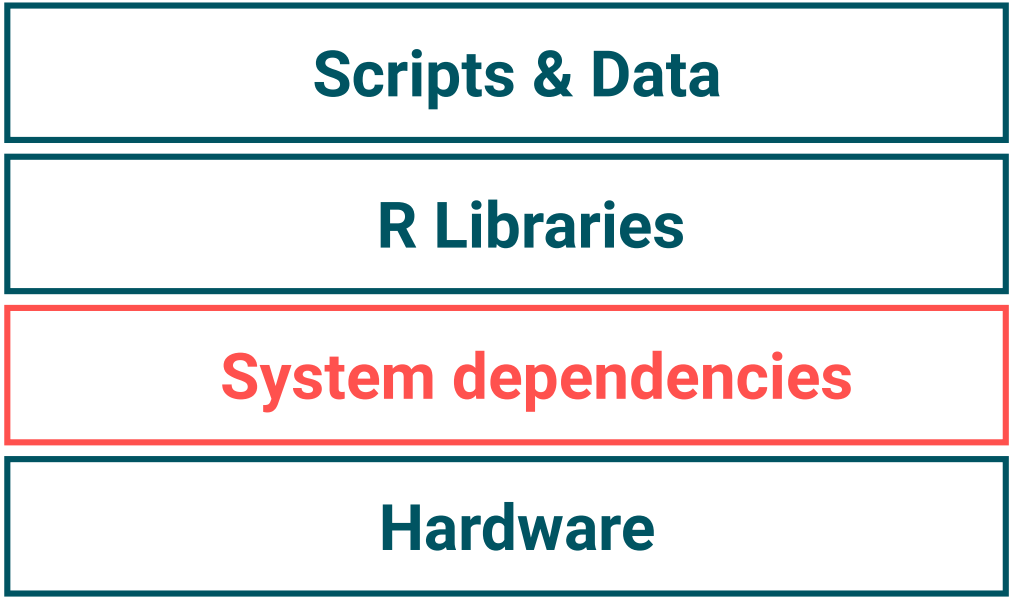
Scripts & Data



```
Warning message:
"`funs()` was deprecated in dplyr 0.8.0.
i Please use a list of either functions or lambdas:

# Simple named list: list(mean = mean, median = median)

# Auto named with `tibble::lst()`: tibble::lst(mean, median)

# Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE
i The deprecated feature was likely used in the dataiku package.
  Please report the issue to the authors."
```

The computational reproducibility stack

Ensure everyone can run your analysis



Scripts & Data

R Libraries

System dependencies

Hardware

git zenodo

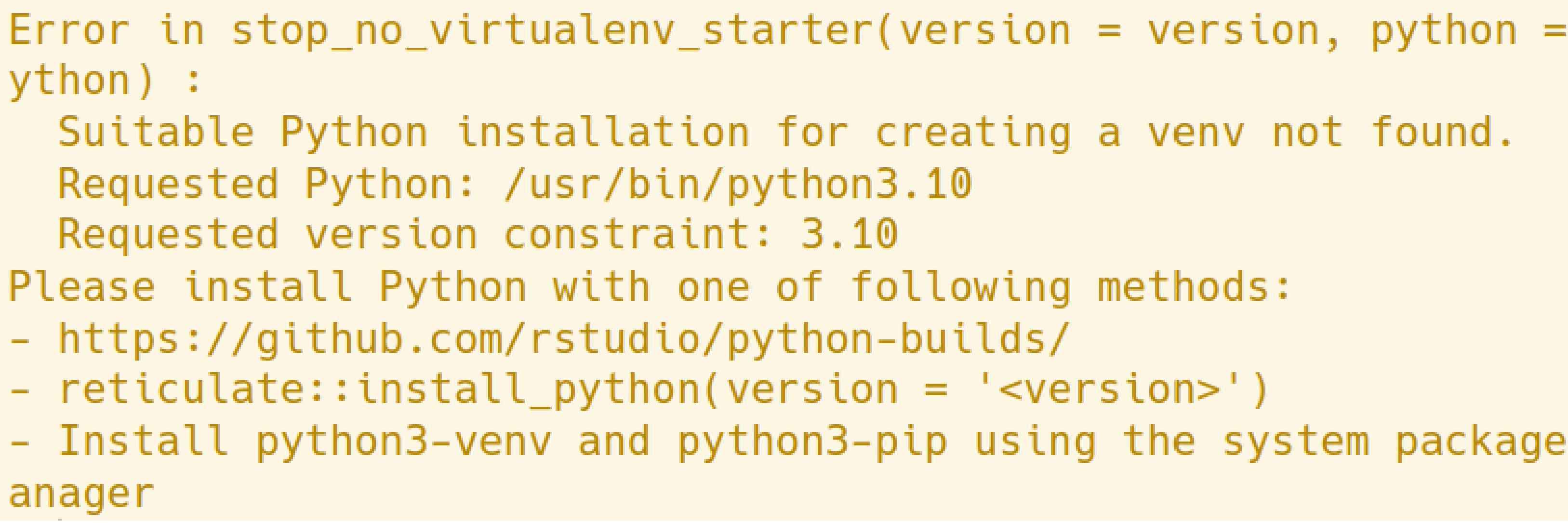
renv

docker

The computational reproducibility stack

Ensure everyone can run your analysis

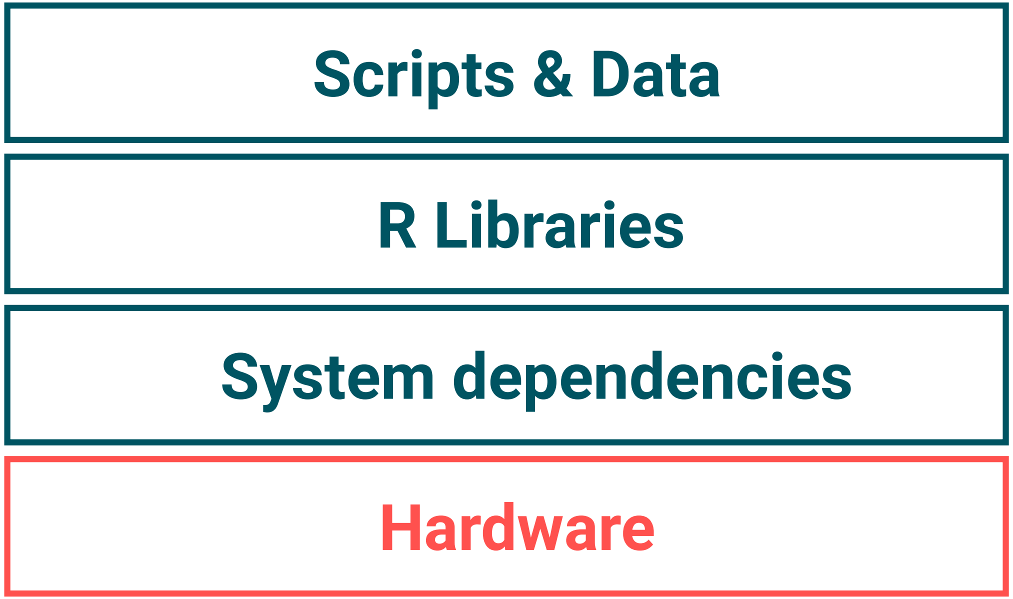
Scripts & Data



```
Error in stop_no_virtualenv_starter(version = version, python =
ython) :
  Suitable Python installation for creating a venv not found.
  Requested Python: /usr/bin/python3.10
  Requested version constraint: 3.10
Please install Python with one of following methods:
- https://github.com/rstudio/python-builds/
- reticulate::install_python(version = '<version>')
- Install python3-venv and python3-pip using the system package
anager
```

The computational reproducibility stack

Ensure everyone can run your analysis



Scripts & Data

R Libraries

System dependencies

Hardware

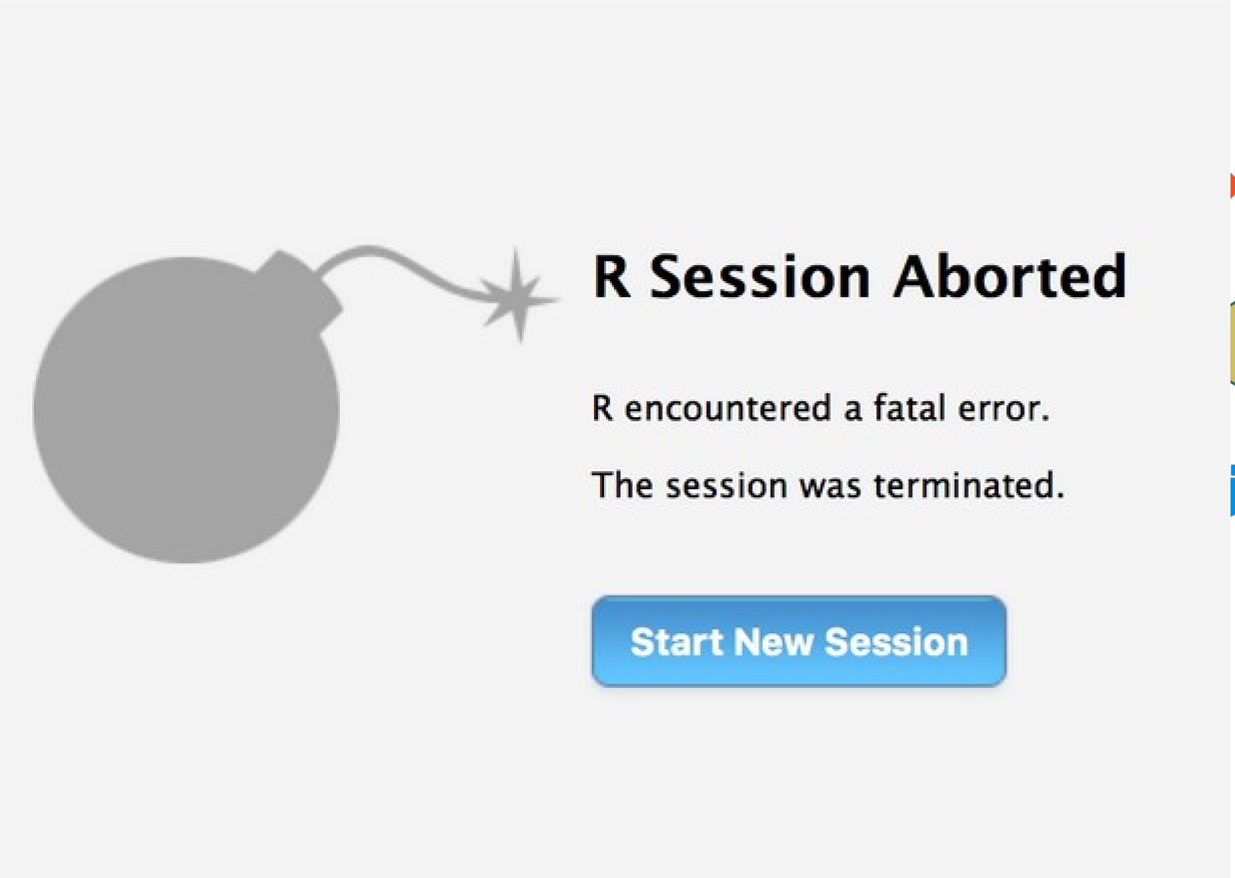
git zenodo

renv

docker

The computational reproducibility stack

Ensure everyo
can run you
analysis



R Session Aborted

R encountered a fatal error.

The session was terminated.

git zenodo

renv

docker

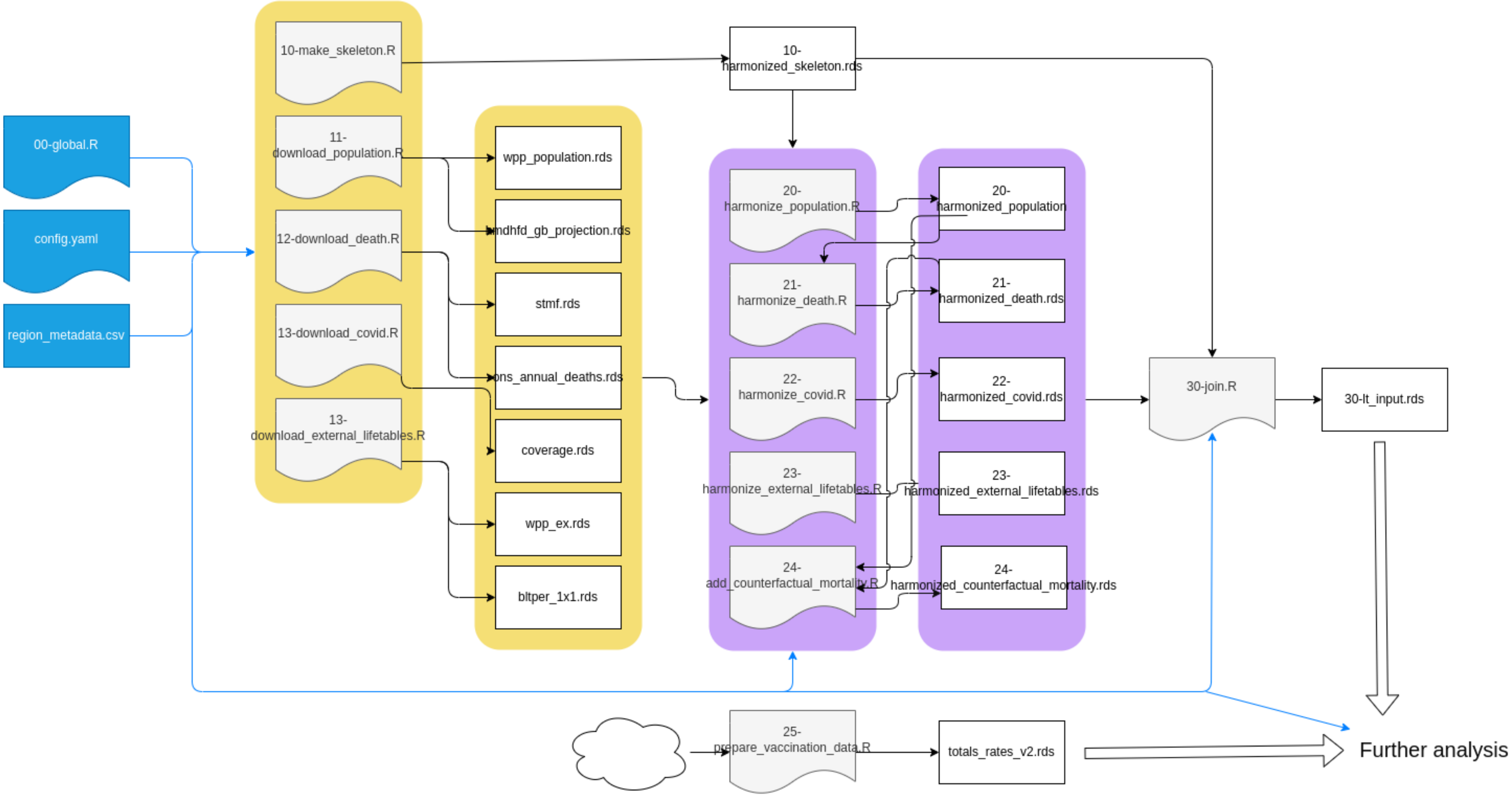
Demonstrating the reproducible workflow

Sharing a statistical model

Demonstrating the reproducible workflow

Life expectancy changes in 2021

The dependency graph



Roadblocks to open science

I can’t share my data

Roadblocks to open science

My code sucks

Roadblocks to open science

Others will copy my stuff

Roadblocks to open science

My co-authors are not on-board

Reproducible analysis

github.com/jschoeley

Jonas Schöley

@jschoeley

0000-0002-3340-8518

schoeley@demogr.mpg.de

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH