

eurostat: R tools for eurostat open data

Leo Lahti, Janne Huovari, Markus Kainu, Przemyslaw Biecek

Abstract An abstract of less than 150 words.

Introductory section which may include references in parentheses (R Core Team, 2013).

Governmental institutions have started to release increasing amount of their data resources for the public as open data in the recent years. This is opening novel opportunities for research and citizen science. Eurostat, for instance, provides a rich collection of demographic and economic data through its open data portal. The portal currently includes ... data sets on ... between years ...

Efficient tools to access and analyse such data collections can greatly benefit reproducible research. When the data is available, the analytical methodology spanning from raw data to the final publication can also be made available following reproducible research principles (Gandrud, 2013). Standardization and automatization of common data analysis tasks via dedicated software packages can greatly facilitate reproducibility and code sharing, making the research more transparent and efficient.

Here, we introduce the eurostat R package that implements automated tools for R to access open data from Eurostat¹ in R (R Core Team, 2013). The package has been actively developed by several independent contributors and based on the community feedback in Github, and its first CRAN release was on.. We are now reporting the first mature version of the package that has been improved and tested by multiple users, and includes features like cache, handling dates, and using the tidy data principles (Wickham, 2014), with the help of the tidyr R package (Wickham, 2015c).

Despite many efforts to this direction, a dedicated R package for eurostat open data has been missing. Our work extends our earlier CRAN packages statfi (Lahti et al., 2013) and smarterpoland (Biecek, 2015). Compared to this earlier work, we have now implemented an expanded set of tools specifically focusing on the eurostat data collection. The datamart (Weinert, 2014) and the quandl (Raymond McTaggart et al., 2015) R packages provide also access to certain versions of eurostat data. In contrast to these generic database packages, our work is fully focused on the Eurostat open data portal and provides specific functionality suited for this data collection. There is a development version for R package reurostat² but it does not seem to be actively maintained at the moment. The eurostat R package takes advantage of the following external R packages: devtools (Wickham and Chang, 2015), dplyr (Wickham and Francois, 2015), knitr (Xie, 2015), ggplot2 (Wickham, 2009), mapproj (for R by Ray Brownrigg et al., 2015), plotrix (J, 2006), reshape2 (Wickham, 2007), rmarkdown (Allaire et al., 2015), stringi (Gagolewski and Tartanus, 2015), testthat (Wickham, 2011), tidyr (Wickham, 2015c). The eurostat R package is part of the rOpenGov project (Leo Lahti and Kainu, 2013), which provides reproducible research tools for computational social science and digital humanities.

Overview of the functionality

The package includes tools to search and retrieve specific data sets from the Eurostat open data portal, converting identifiers in human-readable formats, selecting, modifying and visualizing the data. Further examples are provided in the package vignette³, and a blog post⁴.

The unified interface to data sets can make data analysis more straightforward and transparent by providing a standardized and automated way to access the data sets. Here we describe the functionality of the current CRAN release version (1.2.1). To install this, simply use

```
> install.packages("eurostat")
```

You can download the complete table of contents of the database with the function `'get_eurostat_toc()'`, or use `'search_eurostat()'` to make a more focused search over the table of contents. To retrieve data sets for `'disposable income'`, for instance, use:

```
> library(eurostat)
> income <- search_eurostat("disposable income", type = "dataset")
```

The data type to search for is also specified. The options include `'table'`, `'dataset'` or `'folder'`, referring to different levels of hierarchy in data organization: a `'table'` resides in `'dataset'` that are stored in a `'folder'`. You can focus the search on a selected type.

¹<http://ec.europa.eu/eurostat>

²<https://github.com/Tungurahua/reurostat>

³https://github.com/rOpenGov/eurostat/vignette/eurostat_tutorial.Rmd

⁴<http://ropengov.github.io/r/2015/05/01/eurostat-package-examples>

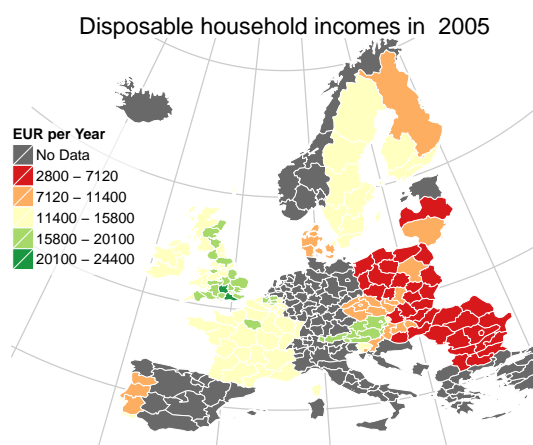


Figure 1: Test caption.

The first lines of the output are shown above. Use values from the ‘code’ column to refer to a specific data set in the subsequent download commands. The dataset identifier codes can also be browsed at the Eurostat database⁵, which gives codes in the Data Navigation Tree after every dataset in parenthesis.

Download the selected data with ‘get_eurostat()’. To retrieve the data set with a particular identifier, use

```
> dat <- get_eurostat(id, time_format = "num")
> print(xtable(head(dat)))
```

	unit	vehicle	geo	time	values
1	PC	BUS_TOT	AT	1990.00	11.00
2	PC	BUS_TOT	BE	1990.00	10.60
3	PC	BUS_TOT	BG	1990.00	
4	PC	BUS_TOT	CH	1990.00	3.70
5	PC	BUS_TOT	CY	1990.00	
6	PC	BUS_TOT	CZ	1990.00	

The original data is annual in this example, hence we have here selected a numeric time variable as it is more convenient for annual time series than the default date format. To improve the interpretability of the output, the eurostat variable identifiers could be further replaced with human-readable labels based on definitions from Eurostat dictionaries with the ‘label_eurostat()’ function. The data is provided in the standard data.frame format, and all standard tools for data subsetting and reshaping can be conveniently applied.

The downloaded data sets are stored in cache by default. This can help to avoid repeated downloading of identical data and helps to speed up the analysis. Another advantage is that by storing an exact copy of the data on the hard disk, it is possible to reproduce the analysis results afterwards even if the source database has been updated.

Data sets containing geographic information can be visualized on a map. Most indicators are of country-year -type, although some indicators have data also at lower level of regional breakdown

The disposable income of private households at NUTS2⁶ level can be visualized, for instance, as in Figure 1. For a more detailed treatment of this example, see our related blog post⁷.

Here, we have combined the data retrieved with the eurostat package with additional map visualization tools and utilities including grid (R Core Team, 2015), maptools (Bivand and Lewin-Koh, 2015), rgdal (Bivand et al., 2015), rgeos (Bivand and Rundel, 2015), scales (Wickham, 2015a), and stringr (Wickham, 2015b).

⁵<http://ec.europa.eu/eurostat/data/database>

⁶http://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics

⁷<http://ropengov.github.io/r/2015/05/01/eurostat-package-examples>

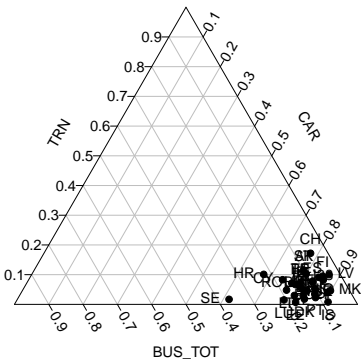


Figure 2: Test caption2.

Example on spatio-temporal data visualization: let us look at the indicator `tgs00026`⁸, (Disposable income of private households by NUTS 2 regions) from Eurostat. We are looking at the disposable household income. In addition to downloading and manipulating data from EUROSTAT, we demonstrate how to access and use shapefiles of Europe published by EUROSTAT at Administrative units / Statistical units⁹.

Further examples

To make a triangle map from the `plotrix` (J, 2006) package provides an example on visualizing passenger transport data distributions (Figure 2):

To facilitate fast plotting of standard European geographic areas, the package provides ready-made lists of the country codes used in the eurostat database for EFTA (`efta_countries`), Euro area (`ea_countries`), EU (`eu_countries`) and EU candidate countries (`candidate_countries`). This helps to select specific groups of countries for closer investigation. For conversions with other standard country coding systems, see the `countrycode` R package (Arel-Bundock, 2014). To retrieve the eurostat country code list for EFTA, for instance, use:

```
> data(efta_countries)
> print(xtable(efta_countries))
```

	code	name
1	IS	Iceland
2	LI	Liechtenstein
3	NO	Norway
4	CH	Switzerland

Applications

The package or its predecessors have already been applied in several case studies by us and independent developers. Financial Times, for instance, have used R to access data from Eurostat¹⁰ using functions from `SmarterPoland`, the direct predecessor of our revised and expanded eurostat package.

The `archivist` R package¹¹ for archivisation of objects has exemplified¹² its functionality by using

⁸<http://ec.europa.eu/eurostat/en/web/products-datasets/-/TGS00026>
⁹<http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units>
¹⁰<http://blog.revolutionanalytics.com/2015/04/financial-times-tracks-unemployment-with-r.html>
¹¹<http://pbiemek.github.io/archivist>
¹²<http://pbiemek.github.io/archivist/justGetIT.html>

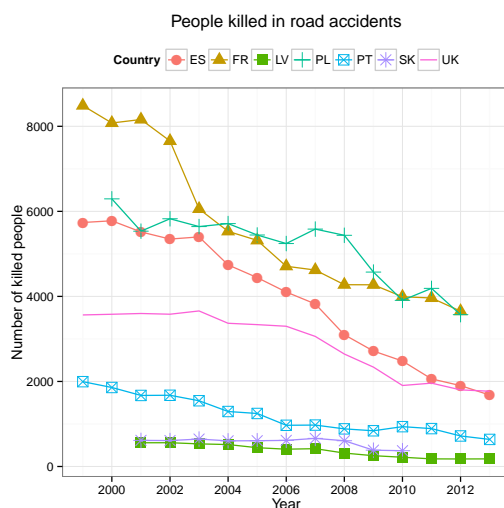


Figure 3: Road accidents caption

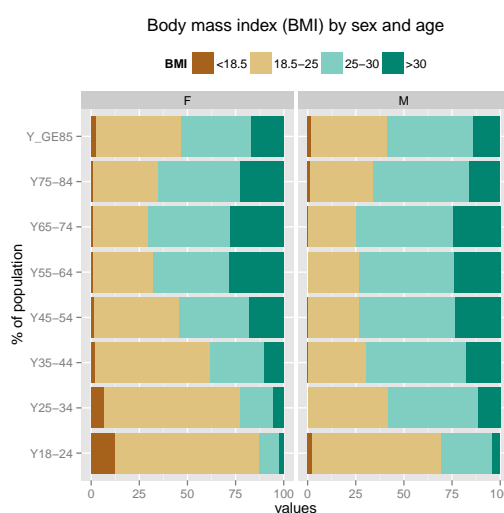


Figure 4: BMI caption

eurostat to plot the number of people killed by road accidents, showing a decreasing trend of road accidents in many countries (Figure 3).

We can also look at the distribution of BMI between different age groups (Figure 4).

Summary

The eurostat R package provides convenient tools to access open data from Eurostat. When automated access to the data sets is integrated with data analytical tools from other packages, this allows a seamless automation of the data analytical process from raw data access to statistical analysis and the final publication.

The package source code can be freely used, modified and distributed under the BSD-2-clause (modified FreeBSD) license. A reproducible version of this article is available at [..](#) and can be used to generate the manuscript text along with up-to-date figures and tables with the latest version of the eurostat data. Manuscript automation provides transparent documentation with full algorithmic details on how to access, preprocess, analyse, and report data and analyses, thus serving a template for good reproducible research practice. The reproducible source code for this manuscript is available

at eurostat github page¹³.

The package exemplifies also the challenges and possible solutions to reproducible research and automated open data retrieval. Possible future extensions and improvements include design of specific data representation class structures. This could facilitate harmonization of the data representation with similar governmental data sets and subsequent tool development. The latest development version of the package can be installed from Github by following the instructions at the github site¹⁴. We welcome issues, bug reports and other feedback via the development site¹⁵.

Acknowledgements

We are grateful to Eurostat¹⁶ for maintaining the open data portal and the rOpenGov¹⁷ for supporting package development. This work has been partially funded by Academy of Finland (decision 293316). We also wish to thank Juuso Parkkinen and Joona Lehtomaki for their feedback on this work.

Bibliography

- J. Allaire, J. Cheng, Y. Xie, J. McPherson, W. Chang, J. Allen, H. Wickham, A. Atkins, and R. Hyndman. *rmarkdown: Dynamic Documents for R*, 2015. URL <http://rmarkdown.rstudio.com>. R package version 0.8.1. [p1]
- V. Arel-Bundock. *countrycode: Convert Country Names and Country Codes*, 2014. URL <http://CRAN.R-project.org/package=countrycode>. R package version 0.18. [p3]
- P. Biecek. *SmarterPoland: Tools for Accessing Various Datasets Developed by the Foundation SmarterPoland.pl*, 2015. URL <http://CRAN.R-project.org/package=SmarterPoland>. R package version 1.5. [p1]
- R. Bivand and N. Lewin-Koh. *maptools: Tools for Reading and Handling Spatial Objects*, 2015. URL <http://CRAN.R-project.org/package=maptools>. R package version 0.8-37. [p2]
- R. Bivand and C. Rundel. *rgeos: Interface to Geometry Engine - Open Source (GEOS)*, 2015. URL <http://CRAN.R-project.org/package=rgeos>. R package version 0.3-14. [p2]
- R. Bivand, T. Keitt, and B. Rowlingson. *rgdal: Bindings for the Geospatial Data Abstraction Library*, 2015. URL <http://CRAN.R-project.org/package=rgdal>. R package version 1.0-7. [p2]
- D. M. P. for R by Ray Brownrigg, T. P. Minka, and transition to Plan 9 codebase by Roger Bivand. *mapproj: Map Projections*, 2015. URL <http://CRAN.R-project.org/package=mapproj>. R package version 1.2-4. [p1]
- M. Gagolewski and B. Tartanus. *R package stringi: Character string processing facilities*, 2015. URL <http://stringi.rexamine.com/>. [p1]
- C. Gandrud. *Reproducible Research with R and R Studio*. Chapman & Hall/CRC, July 2013. [p1]
- L. J. Plotrix: a package in the red light district of r. *R-News*, 6(4):8–12, 2006. [p1, 3]
- L. Lahti, J. Parkkinen, and J. Lehtomaki. *statfi* r package, 2013. [p1]
- J. L. Leo Lahti, Juuso Parkkinen and M. Kainu. ropengov: open source ecosystem for computational social sciences and digital humanities. Presentation at ICML/MLOSS workshop (Int'l Conf. on Machine Learning - Open Source Software workshop), December 2013. URL <http://ropengov.github.io>. [p1]
- R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN ISBN 3-900051-07-0. URL <http://www.R-project.org/>. [p1]
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL <https://www.R-project.org/>. [p2]

¹³<https://github.com/rOpenGov/eurostat/blob/master/vignettes/manuscript.Rmd>

¹⁴<https://github.com/rOpenGov/eurostat>

¹⁵<https://github.com/ropengov/eurostat>

¹⁶<http://ec.europa.eu/eurostat>

¹⁷<https://github.com/ropengov.io>

- Raymond McTaggart, Gergely Daroczi, and Clement Leung. *Quandl: API Wrapper for Quandl.com*, 2015. URL <http://CRAN.R-project.org/package=Quandl>. R package version 2.7.0. [p1]
- K. Weinert. *datamart: Unified access to your data sources*, 2014. URL <http://CRAN.R-project.org/package=datamart>. R package version 0.5.2. [p1]
- H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL <http://www.jstatsoft.org/v21/i12/>. [p1]
- H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL <http://had.co.nz/ggplot2/book>. [p1]
- H. Wickham. testthat: Get started with testing. *The R Journal*, 3:5–10, 2011. URL http://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf. [p1]
- H. Wickham. Tidy data. *Journal of Statistical Software*, 59(10), 2014. [p1]
- H. Wickham. *scales: Scale Functions for Visualization*, 2015a. URL <http://CRAN.R-project.org/package=scales>. R package version 0.3.0. [p2]
- H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2015b. URL <http://CRAN.R-project.org/package=stringr>. R package version 1.0.0. [p2]
- H. Wickham. *tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions*, 2015c. URL <http://CRAN.R-project.org/package=tidyr>. R package version 0.3.1. [p1]
- H. Wickham and W. Chang. *devtools: Tools to Make Developing R Packages Easier*, 2015. URL <http://CRAN.R-project.org/package=devtools>. R package version 1.9.1. [p1]
- H. Wickham and R. Francois. *dplyr: A Grammar of Data Manipulation*, 2015. URL <http://CRAN.R-project.org/package=dplyr>. R package version 0.4.3. [p1]
- Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2015. URL <http://yihui.name/knitr/>. R package version 1.11. [p1]

Leo Lahti

Department of Mathematics and Statistics

PO Box 20014 University of Turku

Finland

leo.lahti@iki.fi

Janne Huovari

Affiliation

Address

Country

author2@work

Markus Kainu

Affiliation

Address

Country

author3@work

Przemysław Biecek

Faculty of Mathematics, Informatics, and Mechanics

University of Warsaw

Banacha 2, 02-097 Warsaw

Poland

P.Biecek@mimuw.edu.pl