eurostat: Eurostat Open Data R Tools DRAFT VERSION IN PROGRESS

Leo Lahti, Janne Huovari, Markus Kainu, Przemyslaw Biecek

Abstract 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. The eurostat R package provides a suite of tools to access open data from Eurostat, including functions to search, download, and manipulate these data sets in an automated and reproducible manner. The online documentation provides further examples on how to visualize, summarize and interpret these spatio-temporal data sets. The package builds on and extends the preceding SmarterPoland and statfi packages and has been extensively tested by the user community. This package contributes to the growing ecosystem of R packages that provide generic tools for reproducible computational research in social science and humanities.

Efficient tools to access and analyse data collections from the public domain can greatly benefit reproducible research (Gandrud, 2013; Boettiger et al., 2015). When the data is available, the complete analytical workflow spanning from raw data to the final publication can be made available. Standardization and automatization of common data analysis tasks via dedicated software packages can help to automate the analysis workflow, thus greatly facilitate reproducibility and code sharing and making data analysis more transparent and efficient.

Here, we introduce the eurostat R package that implements R tools to access open data from Eurostat [1]. Eurostat provides a rich collection of demographic and economic data through its open data portal. The portal currently includes ... data sets on ... between years ...

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 eurostat R package hence brings together earlier, independent efforts by the package developers. It has been actively developed by several contributors and based on the community feedback in Github, with its first CRAN release in 2014. 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). The package or its predecessors have been applied in several case studies by us and independent users including Financial Times[2].

The datamart (Weinert, 2014) and the quandl (Raymond McTaggart et al., 2015) R packages provide generic tools that can be used to access certain versions of eurostat data. In contrast to these generic database packages, our the eurostat package provides functionality that is particularly tailored for this data collection. There is also a development version for R package **reurostat**[3] but this does not seem to be actively maintained. The package depends on or imports 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), and **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.

The package provides tools to search and retrieve data from the Eurostat open data portal, to convert identifiers in human-readable formats and to select, modify and visualize the data. In this manuscript, we provide a brief overview of the core functionality in the current CRAN release version (1.2.1). For further examples, see the package vignette[4].

Search and download commands

To install the CRAN release version, type in R:

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

The complete table of contents of the database can be downloaded in R with the function `get_eurostat_toc()` [HERE LINK TO EUROSTAT PAGE WHERE THE DATA CAN BE BROWSED

[1] http://ec.europa.eu/eurostat
[2] http://blog.revolutionanalytics.com/2015/04/financial-times-tracks-unemployment-with-r.html
[3] https://github.com/Tungurahua/reurostat
[4] https://github.com/rOpenGov/eurostat/vignette/eurostat_tutorial.Rmd

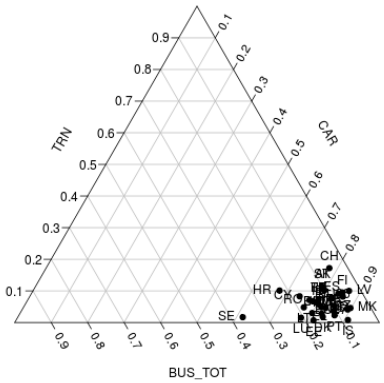

Figure 1: Passenger transport data visualization on a triangular map.

ONLINE]. The function `search_eurostat()` can be used 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 queried data type is specified in the above example with the `type` argument. The options for this argument include *'table'*, *'dataset'* or *'folder'*, referring to different levels of hierarchy in the data organization: a table resides in dataset, and the datasets are stored in a folder. The `type` argument limits the search on a selected data set type. Values from the code column list identifiers for specific data sets. These can be used in subsequent download commands. Alternatively, the dataset identifier codes can also be browsed at the Eurostat website[5]; check the codes in the Data Navigation Tree listed after each dataset (in parentheses).

To retrieve the data set with a particular identifier (tsdtr210, for instance), use

```
dat <-get_eurostat(id = 'tsdtr210',time_format = "num")
```

Since the original data is annual in this example, we have selected a numeric time variable as this is more convient for annual time series than the default date format. The function call returns a table on passenger transport statistics in various countries. The first lines of the output are shown in Table 1.

	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	

Table 1: First lines of output from the `get_eurostat` function for the data set with the identifier 'tsdtr210'.

Multidimensional data sets can be also visualized as triangular maps (Figure 1) based on the **plotrix** (J, 2006) package.

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, so that all standard tools for data subsetting and reshaping are supported.

[5]http://ec.europa.eu/eurostat/data/database

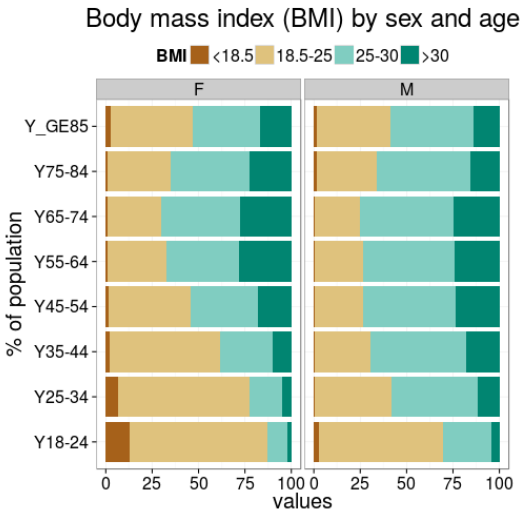

Figure 2: The body-mass index in different age groups based on Eurostat open data.

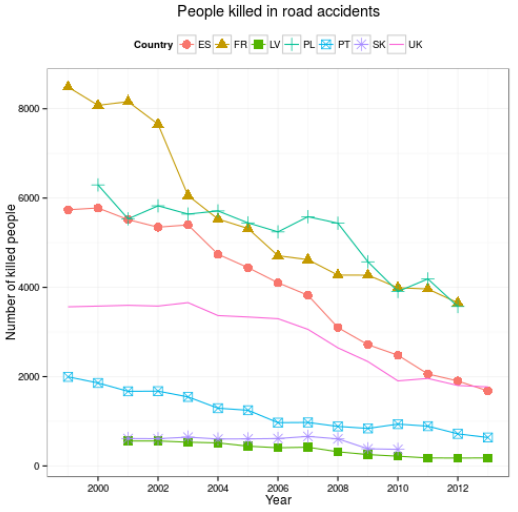

Figure 3: Time series of the number of people killed in road accidents.

Usage examples

Eurostat provides a number of data sets of demographics and economics. We can look, for instance, at the distribution of body-mass index (BMI) indicating differences in obesity between different age groups as visualized in Figure 2, or observe a decreasing trend of road accidents in many countries over time (Figure 3)[6].

Geospatial information

Spatial information is conveniently visualized on a map. Most indicators are of country/year-type, although sometimes data is available also at more detailed levels of regional breakdown. Our related blog post[7] provides a detailed description on geospatial data visualization, looking at disposable income of private households (eurostat id tgs00026[8]) at the NUTS2 regions[9] (Figure 4).

[6] http://pbiecek.github.io/archivist/justGetIT.html
[7] http://ropengov.github.io/r/2015/05/01/eurostat-package-examples
[8] http://ec.europa.eu/eurostat/en/web/products-datasets/-/TGS00026
[9] http://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics

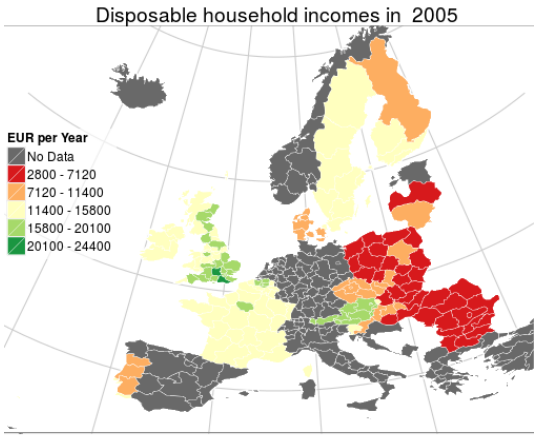

Figure 4: Disposable income of private households shown on the European map retrieved with the eurostat package and visualized with the aid of additional R extensions.

In addition to downloading and manipulating data from Eurostat, we demonstrate how to access and use the Eurostat geospatial data (shapefiles) on administrative statistical units[10] to visualize the income statistics on the European map. The example demonstrates how the Eurostat data sets and geospatial data (shapefiles), retrieved with the eurostat package, can be combined with additional map visualization tools and other 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).

Other functionality

To facilitate convenient analysis and visualization of standard European areas, we have included ready-made country code lists for certain standard areas. To retrieve, for instance, the eurostat country code list for EFTA as shown in Table **??**, use:

```
data(efta_countries)
```

Such country listings are available for EFTA (efta_countries), Euro area (ea_countries), EU (eu_countries) and EU candidate countries (candidate_countries) [MENTION WHICH COUNTRY CODING SYSTEM IS BEING USED HERE]. These auxiliary data sets facilitate fast selection of specific country groups. Conversions between different country coding standards are available with the **countrycode** R package (Arel-Bundock, 2014).

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

Table 2: The EFTA country code table as provided by the eurostat R package.

The downloaded data sets are stored in cache by default to avoid repeated downloads of identical data sets, helping to speed up the analysis. Another advantage is that an exact copy of the retrieved data on the hard disk makes it possible to reproduce the analysis results even when the source database is updated.

[10] http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units

Summary

The eurostat R package provides convenient tools to access open data from Eurostat. Integrating automated access to the data sets with further data analysis and visualization tools provided in other packages allows a seamless automation of the data analytical workflow from accessing the raw data to statistical analysis and final publication. For the full reproducible source code of the figures and tables of this manuscript, see the package Github site[11]. The reproducible Rmarkdown document provides transparent documentation with full algorithmic details on how to access, preprocess, analyse, and report data and analyses, and can be also used to update the material when new or updated versions of the Eurostat data become available.

The package provides one set of solutions to automated data retrieval from independent data repositories, featuring options such as search, subsetting and cache. Possible future extensions and improvements include implementation of specific data representation formats to harmonize the data representation [LINK TO PANEL-SERIES DATA HERE] across similar data sources and to facilitate subsequent tool development.

The latest development version of the package can be installed from Github by following the instructions at the github site[12]. The package source code can be freely used, modified and distributed under the BSD-2-clause (modified FreeBSD) license. We welcome issues, bug reports and other feedback via the development site[13].

Acknowledgements

We are grateful to Eurostat[14] for maintaining the open data portal and the rOpenGov[15] 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. [p4]

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. [p4]

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. [p4]

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. [p4]

C. Boettiger, S. Chamberlain, E. Hart, and K. Ram. Building software, building community: Lessons from the ropensci project. *Journal of Open Research Software*, 3(1), November 2015. doi: http://doi.org/10.5334/jors.bu. [p1]

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]

[11] https://github.com/rOpenGov/eurostat/blob/master/vignettes/manuscript.Rmd
[12] https://github.com/rOpenGov/eurostat
[13] https://github.com/ropengov/eurostat
[14] http://ec.europa.eu/eurostat
[15] https://github.ropengov.io

L. J. Plotrix: a package in the red light district of r. *R-News*, 6(4):8–12, 2006. [p1, 2]

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, 2015. URL https://www.R-project.org/. [p4]

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. [p4]

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. [p4]

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

Przemyslaw Biecek
Faculty of Mathematics, Informatics, and Mechanics

University of Warsaw
Banacha 2, 02-097 Warsaw
Poland
P.Biecek@mimuw.edu.pl