eurostat: Eurostat Open Data R Tools DRAFT VERSION IN PROGRESS

Leo Lahti, Janne Huovari, Markus Kainu, Przemyslaw Biecek

Abstract Whereas open data released by governmental institutions are opening up novel opportunities for research and citizen science, efficient tools to access and analyze these data sets are needed to realize the full potential of these information resources. We introduce here the **eurostat** R package that provides a suite of tools to access open data from Eurostat, including functions to search, download, and manipulate Eurostat data in an automated and reproducible manner. The online documentation provides detailed examples on how to access, summarize and visualize these spatio-temporal data sets. The package expands previous related work and has been extensively tested by the user community. This contributes to the growing ecosystem of R packages that provide algorithmic tools for reproducible computational research in social science and humanities.

Eurostat[1], the statistical office of the European Union, is providing a rich collection of European level demographic and economic data through its open data service, which currently includes over 8800 data sets on European demography, economics, health, infrastructure, traffic and other topics. In many cases the statistics are available with great geographical resolution and as time series spanning over several years or decades.

The availability of 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 resources and analysis algorithms are openly available, the complete analytical workflow spanning from raw data to the final publication can be made fully automated and transparent. Standardization of common data analysis tasks via dedicated software packages can help to automate the analysis workflow, greatly facilitating reproducibility and code sharing, and making the data analysis more efficient. At the same time, the algorithms need to be customized for variations in data formats, access details, and typical use cases.

Here, we introduce the **eurostat** R package that implements R tools specifically designed to facilitate such automated access to open data from Eurostat [2]. Despite earlier efforts, a dedicated R package for eurostat open data has been missing. The eurostat R package introduced here brings together earlier efforts from our earlier CRAN packages **statfi** (Lahti et al., 2013) and **smarterpoland** (Biecek, 2015). Compared to this earlier work, we have combined the relevant parts of these two packages and implemented an expanded set of tools with a specific focus on the Eurostat data collection. The first version of the new **eurostat** package was released in CRAN in 2014. It has been actively developed by several contributors and based on community feedback in Github. We are now reporting the first mature version of the package that has been improved and tested by multiple users. The package and its predecessors have been applied in several case studies by us and others[3].

Related work includes the **datamart** (Weinert, 2014) and the **quandl** (Raymond McTaggart et al., 2015) R packages that provide generic tools that can be used to access certain versions of Eurostat data. In contrast to these generic database packages, our eurostat package provides functionality that is particularly tailored for the Eurostat open data service. The development version of another related R package **reurostat**[4] does not seem to be actively maintained at the moment. Moreover, our **eurostat** package depends, imports or suggests 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 rOpenGov collection (Leo Lahti and Kainu, 2013) that provides reproducible research tools for computational social science and digital humanities.

In summary, the **eurostat** package provides custom tools to search, retrieve, modify and visualize data from the Eurostat open data service. The package supports key features such as data cache, date formatting, and tidy data principles (Wickham, 2014) using the the **tidyr** R package (Wickham, 2015c). Here, we provide an overview of the core functionality in the current CRAN release version (1.2.1). For further examples, see the package vignette[5].

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

Search and download commands

To install and load the CRAN release version, just type in R:

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

The complete table of contents of the database can be browsed on-line[6], or downloaded in R with the command `toc <-get_eurostat_toc()`. The function `search_eurostat()` is used to make a more focused search over the table of contents. To retrieve data for 'Modal split of passenger transport', for instance, use:

```
> query <- search_eurostat("Modal split of passenger transport", type = "table")
```

The `type` argument limits the search on a selected data set type in the above example. 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`, which is in turn stored in a `folder`.

Values in the `code` column of the `search_eurostat()` function output provide data sets identifiers that can be used in subsequent download commands. Alternatively, these identifier codes can be browsed at the Eurostat open data service; check the codes in the Data Navigation Tree listed after each dataset in parentheses. Let us look at the data set identifier and title for the first entry of the query data:

```
> query$code[[1]]
[1] "tsdtr210"

> query$title[[1]]
[1] "Modal split of passenger transport"
```

To retrieve the data set with this identifier, use

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

As the original data is annual in this example, we have selected a numeric time format. This is more convient for annual time series than the default date format. The data sets are provided as standard data frames to support standard tools for data subsetting and reshaping. The above function call returns a table on transport statistics. 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'.

	unit	vehicle	geo	time	values
1	Percentage	Motor coaches, buses and trolley buses	Austria	1990.00	11.00
2	Percentage	Motor coaches, buses and trolley buses	Belgium	1990.00	10.60
3	Percentage	Motor coaches, buses and trolley buses	Bulgaria	1990.00	
4	Percentage	Motor coaches, buses and trolley buses	Switzerland	1990.00	3.70
5	Percentage	Motor coaches, buses and trolley buses	Cyprus	1990.00	
6	Percentage	Motor coaches, buses and trolley buses	Czech Republic	1990.00	

Table 2: The output from `get_eurostat()` (see Table 1), converted into human-readable labels retrieved by `label_eurostat()`.

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

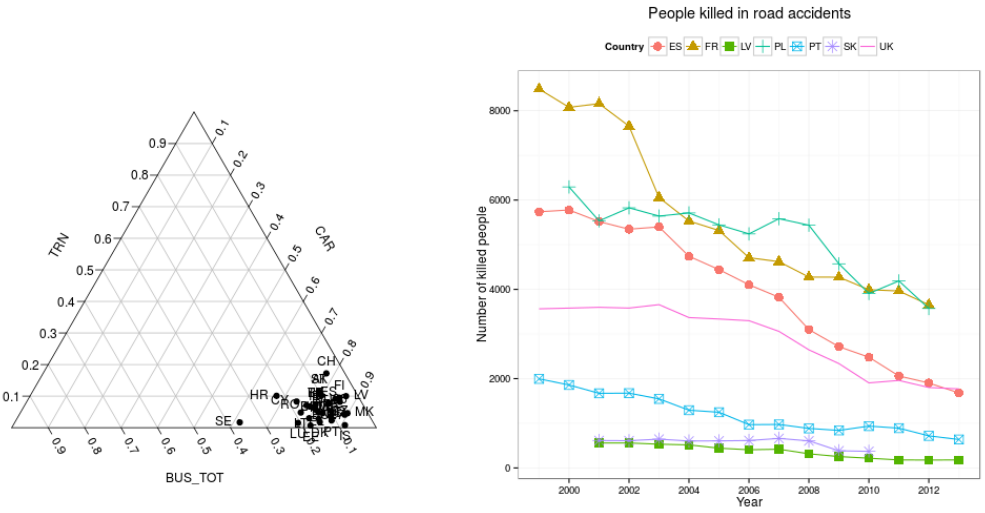

Figure 1: A Passenger transport data retrieved with the **eurostat** package visualized on a triangular **plotrix** map. **B** Time series of the number of people killed in road accidents.

The entries in many columns of the Table 1 are as such not readily interpretable. To improve interpretability, these variable identifiers can be replaced with human-readable labels with the `label_eurostat()` function; the simple call `label_eurostat(dat)` converts the original identifiers into a human-readable version (Table 2) based on translations available from the Eurostat database.

The downloaded data sets are stored in cache by default to avoid repeated downloads of identical data sets. This can speed up the analysis. Moreover, storing an exact copy of the retrieved raw data on the hard disk supports reproducibility when the source database is constantly updated.

Visualizing the Eurostat data

The transport data set in our example includes three classes of vehicles. Three-dimensional data sets such as this can be conveniently visualized as triangular maps by using the **plotrix** (J, 2006) package. The Figure 1A illustrates the distribution of vehicle types in different countries. Interestingly, the Eurostat data also reveals a decreasing trend of road accidents in many countries over time (Figure 1)B as described in more detail in our recent blog post[7].

The Eurostat database includes a variety of demographic and health indicators. We see, for instance, that overweight varies remarkably across different age groups quantified by the body-mass index (BMI) (Figure 2)[8].

Geospatial information

The indicators in the Eurostat open data service are typically available as annual time series grouped by country, and in some cases at more refined temporal or geographic levels. Importantly, Eurostat provides complementary geospatial data on administrative statistical units, available as standard shapefiles[9]. These can be used to visualize the available statistics on the European map at the appropriate geographic resolution.

As an example of geospatial data visualization, let us look at disposable income of private households (data set identifier tgs00026[10]). This data set is available at the geographic level of NUTS2 regions, which is the intermediate level of territorial units in the Eurostat regional classifications, roughly correspond to provinces or states in each country[11] (Figure 3).

[7] http://pbiecek.github.io/archivist/justGetIT.html

[8] The fully reproducible source code of this example is available at https://github.com/rOpenGov/eurostat/blob/master/vignettes/2015-RJournal/lahti-huovari-kainu-biecek.md

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

[10] http://ec.europa.eu/eurostat/en/web/products-datasets/-/TGS00026

[11] http://ec.europa.eu/eurostat/web/nuts/overview

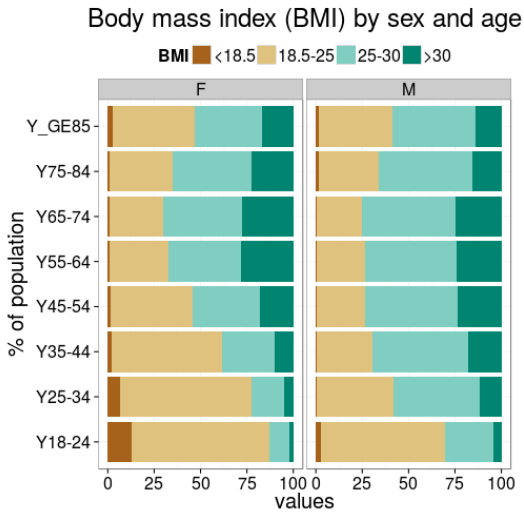

Figure 2: The body-mass index in different age groups based on Eurostat open data.

A detailed treatment of this example, together with reproducible source code, is available on-line[12]. 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).

To further facilitate the 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 codes for EFTA countries (Table ??), we can use:

```
data(efta_countries)
```

Similar country listings are available for EFTA (efta_countries), Euro area (ea_countries), EU (eu_countries) and EU candidate countries (candidate_countries). These auxiliary data sets facilitate fast selection of specific country groups. The full name and a two-letter identifier is provided for each country as available from the Eurostat database. The country codes follow the ISO 3166-1 alpha-2 standard, except that GB and GR are replaced by UK (United Kingdom) and EL (Greece) in the Eurostat database, respectively. Linking these country codes with external data sets can be facilitated by conversions between different country coding standards with the **countrycode** R package (Arel-Bundock, 2014).

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

Table 3: The EFTA country code table from the eurostat R package.

Summary

The eurostat R package provides convenient tools to access open data from Eurostat. Combining programmatic access to the data sets with further analysis and visualization tools allows a seamless and reproducible automation of the complete data analytical workflow from accessing the raw data to statistical analysis and final publication. The full source code of the figures and tables of this manuscript are available at the package Github site[13]. The Rmarkdown document provides transparent

[12] http://ropengov.github.io/r/2015/05/01/eurostat-package-examples

[13] https://github.com/rOpenGov/eurostat/blob/master/vignettes/2015-RJournal/lahti-huovari-kainu-biecek.md

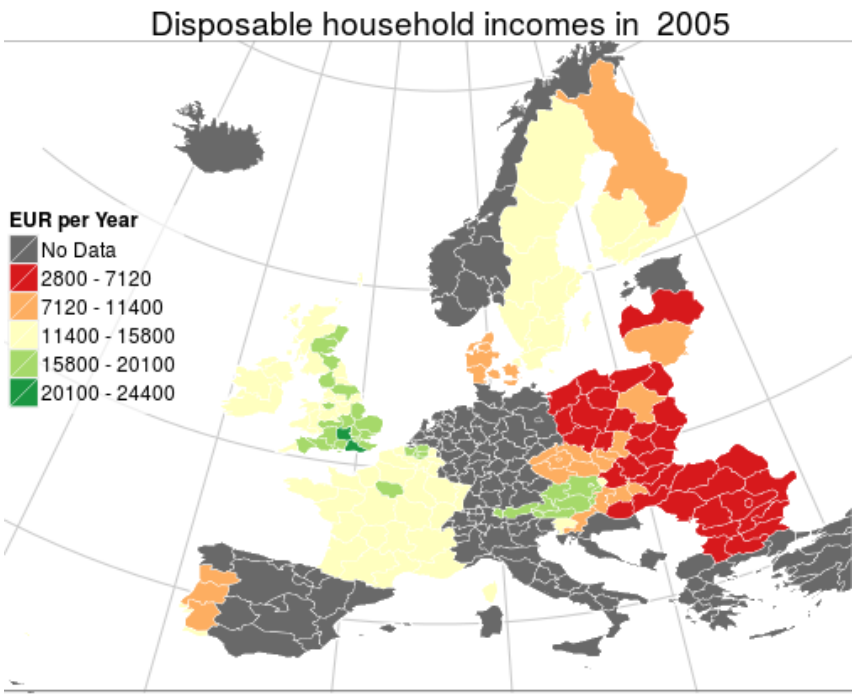

Figure 3: Disposable income of private households across NUTS2-level national regions in European countries visualized based on geospatial data available from Eurostat.

documentation with full algorithmic details on the analyses, and can be updated when new versions of the Eurostat data become available.

The eurostat package provides an example of automated data retrieval from institutional 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 across similar data sources and to facilitate subsequent tool development. In particular, we should take further advantage of the existing spatiotemporal data structures available in R, such as those provided by the **spacetime** package **?**, and construct wrapper functions to speed up routine operations such as visualizing the temporal and geospatial data sets from Eurostat.

To install the latest development version of the eurostat package, follow the instructions at the github site[14]. 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[15].

Acknowledgements

We are grateful to Eurostat[16] for maintaining the open data portal and the rOpenGov[17] 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 Lehtomäki for feedback.

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]

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]

[14]https://github.com/rOpenGov/eurostat
[15]https://github.com/ropengov/eurostat
[16]http://ec.europa.eu/eurostat
[17]https://github.ropengov.io

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