

Retrieval and Analysis of Eurostat Open Data with the `eurostat` Package

by Leo Lahti, Janne Huovari, Markus Kainu, and Przemysław Biecek

Abstract The increasing availability of open statistical data resources is providing novel opportunities for research and citizen science. Efficient algorithmic tools are needed to realize the full potential of the new information resources. We introduce the `eurostat` R package that provides a collection of custom tools for the Eurostat open data service, including functions to query, download, manipulate, and visualize these data sets in a smooth, automated and reproducible manner. The online documentation provides detailed examples on the analysis of these spatio-temporal data collections. This work provides substantial improvements over the previously available tools, and has been extensively tested by an active user community. The `eurostat` R package contributes to the growing open source ecosystem dedicated to reproducible research in computational social science and digital humanities.

Introduction

Eurostat, the statistical office of the European Union, provides a rich collection of data through its open data service¹, including thousands of data sets on European demography, economics, health, infrastructure, traffic and other topics. The statistics are often available with fine geographical resolution and include time series spanning over several years or decades.

Availability of algorithmic tools to access and analyse open data collections can greatly benefit reproducible research (Gandrud, 2013; Boettiger et al., 2015), as complete analytical workflows spanning from raw data to final publications can be made fully replicable and transparent. Dedicated software packages help to simplify, standardize, and automate analysis workflows, greatly facilitating reproducibility, code sharing, and efficient data analytics. The code for data retrieval need to be customized to specific data sources to accommodate variations in raw data formats, access details, and typical use cases so that the end users can avoid repetitive programming tasks and save time. A number of packages for governmental and other sources have been designed to meet these demands, including packages for the Food and Agricultural Organization (FAO) of the United Nations (FAOSTAT; Kao et al. 2015), World Bank (WDI; Arel-Bundock 2013), national statistics authorities (pxweb; Magnusson et al. 2014), Open Street Map (osmar; Eugster and Schlesinger 2012) and many other sources.

A dedicated R package for the Eurostat open data has been missing. The `eurostat` package fills this gap. It expands the capabilities of our earlier `statfi` (Lahti et al., 2013a) and `smarterpoland` (Biecek, 2015) packages. Since its first CRAN release in 2014, the `eurostat` package has been developed by several active contributors based on frequent feedback from the user community. We are now reporting mature version that has been improved and tested by multiple users, and applied in several case studies by us and others (Kenett and Shmueli, 2016). The Eurostat database has three services for programmatic data access: a bulk download, json/unicode, and SDMX web service; we provide targeted methods for the first two in the `eurostat` package; generic tools for the SDMX format are available via the `rsdmx` package (Blondel, 2017). The bulk download provides single files, which is convenient and fast for retrieving major parts of data. More light-weight json methods allow data subsetting before download and may be preferred in more specific retrieval tasks but the query size is limited to 50 categories. Finally, the package can be used to download custom administrative boundaries by EuroGeographics© that allow seamless visualization of the data on the European map.

Specific versions of the Eurostat data can be accessed with the `datamart` (Weinert, 2014), `quandl` (McTaggart et al., 2015), `pdfetch` (Reinhart, 2015), and `rsdmx` packages. Unlike these generic database packages, `eurostat` is particularly tailored for the Eurostat open data service. It depends on further R packages including `classInt` (Bivand, 2015), `httr` (Wickham, 2016), `jsonlite` (Ooms, 2014), `readr` (Wickham and Francois, 2015), `sp` (Pebesma and Bivand, 2005), and `stringi` (Gagolewski and Tartanus, 2015). The `eurostat` package is part of the rOpenGov collection (Lahti et al., 2013b) that provides reproducible research tools for computational social science and digital humanities.

In summary, the `eurostat` package provides custom tools for Eurostat open data. Key features such as cache, date formatting, tidy data principles (Wickham, 2014), and `tibble` (Wickham et al., 2016) data format support seamless integration with other tools for data manipulation and analysis. This article provides an overview of the core functionality in the current CRAN release version (3.1.1). A comprehensive documentation and source code are available via the package homepage in Github².

¹<http://ec.europa.eu/eurostat/data/database>

²<http://ropengov.github.io/eurostat>

Search and download commands

To install and load the CRAN release version, just type the standard installation command in R.

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

The database table of contents is available on-line³, or can be downloaded in R with `get_eurostat_toc()`. A more focused search is provided by the `search_eurostat()` function.

```
query <- search_eurostat("road accidents", type = "table")
```

This seeks data on road accidents. The type argument limits the search on a selected data set type, one of three hierarchical levels including *"table"*, which resides in *"dataset"*, which is in turn stored in a *"folder"*. Values in the code column of the `search_eurostat()` function output provide data identifiers for subsequent download commands. Alternatively, these identifiers 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 identifier and title for the first entry of the query data.

```
query$code[[1]]
[1] "tsdtr420"

query$title[[1]]
[1] "People killed in road accidents"
```

Let us next retrieve the data set with this identifier.

```
dat <- get_eurostat(id = "tsdtr420", time_format = "num")
```

Here we used the numeric time format as it is more convient for annual time series than the default date format. The transport statistics returned by this function call (Table 1) could be filtered before download with the `filters` argument, where the list names and values refer to Eurostat variable and observation codes, respectively. To retrieve transport statistics for specific countries, for instance, use the `get_eurostat` function.

```
countries <- c("UK", "SK", "FR", "PL", "ES", "PT")
t1 <- get_eurostat("tsdtr420", filters = list(geo = countries))
```

	unit	sex	geo	time	values
1	NR	T	AT	1999.00	1079.00
2	NR	T	BE	1999.00	1397.00
3	NR	T	CZ	1999.00	1455.00
4	NR	T	DE	1999.00	7772.00
5	NR	T	DK	1999.00	514.00
6	NR	T	EL	1999.00	2116.00

Table 1: First entries of the road accident data set retrieved with `get_eurostat(id = "tsdtr420", time_format = "num")`.

	unit	sex	geo	time	values
1	Number	Total	Austria	1999.00	1079.00
2	Number	Total	Belgium	1999.00	1397.00
3	Number	Total	Czech Republic	1999.00	1455.00
4	Number	Total	Germany (until 1990 former territory of the FRG)	1999.00	7772.00
5	Number	Total	Denmark	1999.00	514.00
6	Number	Total	Greece	1999.00	2116.00

Table 2: The `get_eurostat()` output (Table 1) converted into human-readable labels with `label_eurostat()`.

A subsequent visualization reveals a decreasing trend of road accidents over time in Figure 1.

³<http://ec.europa.eu/eurostat/data/database>

```
ggplot(t1, aes(x = time, y = values, color = geo, group = geo, shape = geo)) +
  geom_point(size = 4) + geom_line() + theme_bw() +
  ggtitle("Road accidents") + xlab("Year") + ylab("Victims (n)") +
  theme(legend.position = "none") +
  ggrepel::geom_label_repel(data = t1 %>% group_by(geo) %>% na.omit() %>%
    filter(time %in% c(min(time), max(time)))), aes(fill = geo, label = geo), color = "white")
```

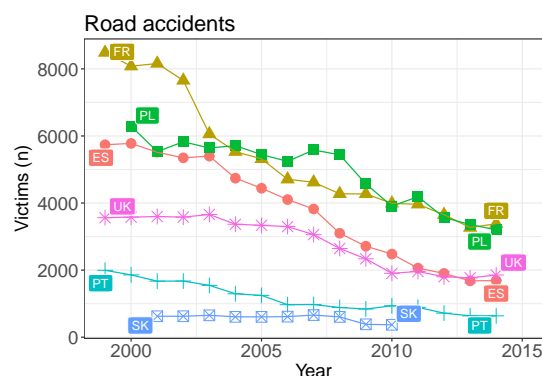


Figure 1: Timeline indicating the number of people killed in road accidents in various countries based on Eurostat open data retrieved with the `eurostat` R package.

Utilities

Many entries in Table 1 are not readily interpretable, but a simple call `label_eurostat(dat)` can be used to convert the original identifier codes into human-readable labels (Table 2) based on translations in the Eurostat database. Labels are available in English, French and German languages.

The Eurostat database includes a variety of demographic and health indicators. We see, for instance, that overweight varies remarkably across different age groups (Figure 2A). Sometimes the data sets require more complicated pre-processing. Let's consider, for instance, the distribution of renewable energy sources in different European countries. In order to summarise such data one needs to first aggregate a multitude of possible energy sources into a smaller number of coherent groups. Then one can use standard R tools to process the data, chop country names, filter countries depending on production levels, normalize the within country production. After a series of transformations (see Appendix for the source code) we can finally plot the data to discover that countries vary a lot in terms of renewable energy sources (Figure 2B). Three-dimensional data sets such as this can be conveniently visualized as triangular maps by using the `plotrix` (Lemon, 2006) package.

The data sets are stored in cache by default to avoid repeated downloads of identical data and to speed up the analysis. Storing an exact copy of the retrieved raw data on the hard disk will also support reproducibility when the source database is constantly updated.

Geospatial information

Map visualizations

The indicators in the Eurostat open data service are typically available as annual time series grouped by country, and sometimes at more refined temporal or geographic levels. Eurostat provides complementary geospatial data on the corresponding administrative statistical units to support visualizations at the appropriate geographic resolution. The geospatial data sets are available as standard shapefiles⁴. Let us look at disposable income of private households (data identifier `tgs00026`⁵). This is provided at the geographic NUTS2 regions, the intermediate territorial units in the Eurostat regional classifications, roughly corresponding to provinces or states in each country⁶ (Figure 3). The map can be generated with the following code chunk.

⁴<http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units>

⁵<http://ec.europa.eu/eurostat/en/web/products-datasets/-/TGS00026>

⁶<http://ec.europa.eu/eurostat/web/nuts/overview>

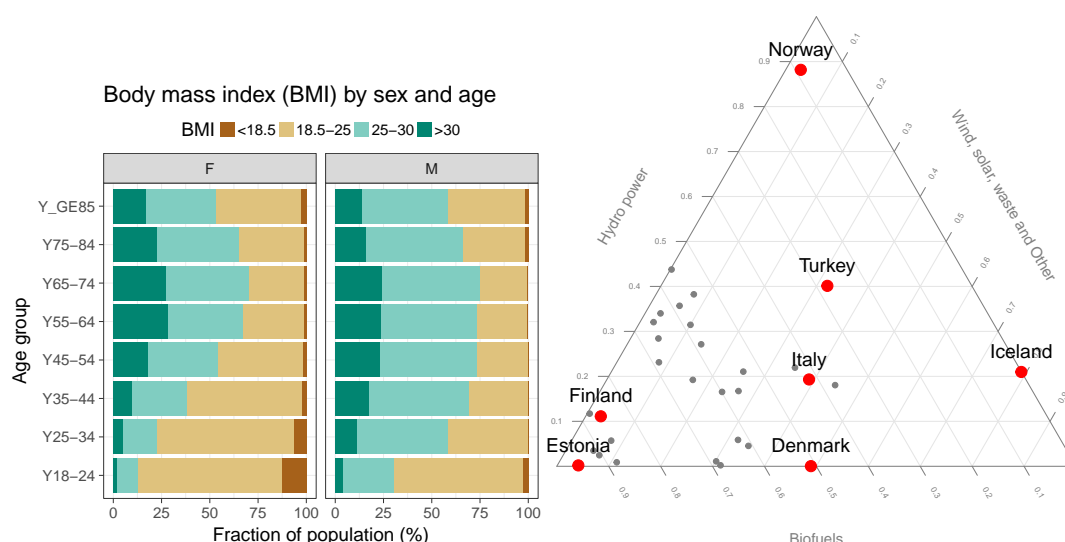


Figure 2: A The body-mass index (BMI) in different age groups in Poland (Eurostat table hlth_ehis_de1). B Production of renewable energy in various countries in 2013 (Eurostat table ten00081). See the Appendix for the source code.

```
# Load the required libraries
library(eurostat)
library(dplyr)
library(ggplot2)

# Download and manipulate tabular data
get_eurostat("tgs00026", time_format = "raw") %>%
  # Subset to year 2005 and NUTS-3 level
  dplyr::filter(time == 2005, nchar(as.character(geo)) == 4) %>%

# Classify the values the variable
dplyr::mutate(cat = cut_to_classes(values)) %>%

# Merge Eurostat data with geodata from Cisco
merge_eurostat_geodata(data = ., geocolumn = "geo", resolution = "60",
  output_class = "df", all_regions = TRUE) %>%

# Plot the map
ggplot(data = ., aes(long, lat, group = group)) +
  geom_polygon(aes(fill = cat), colour = alpha("white", 1/2), size = .2) +
  scale_fill_manual(values = RColorBrewer::brewer.pal(n = 5, name = "Oranges")) +
  labs(title = "Disposable household income") +
  coord_map(project = "orthographic", xlim = c(-22, 34), ylim = c(35, 70)) + theme_minimal() +
  guides(fill = guide_legend(title = "EUR per Year",
    title.position = "top", title.hjust = 0))
```

This demonstrates how the Eurostat statistics and geospatial data, retrieved with the `eurostat` package, can be combined with other utilities, in this case the `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) R packages.

Standard country groupings

To facilitate the analysis and visualization of standard European country groups, the `eurostat` package includes ready-made country code lists. The list of EFTA countries (Table 3), for instance, is retrieved with the `data` command.

```
data(efta_countries)
```

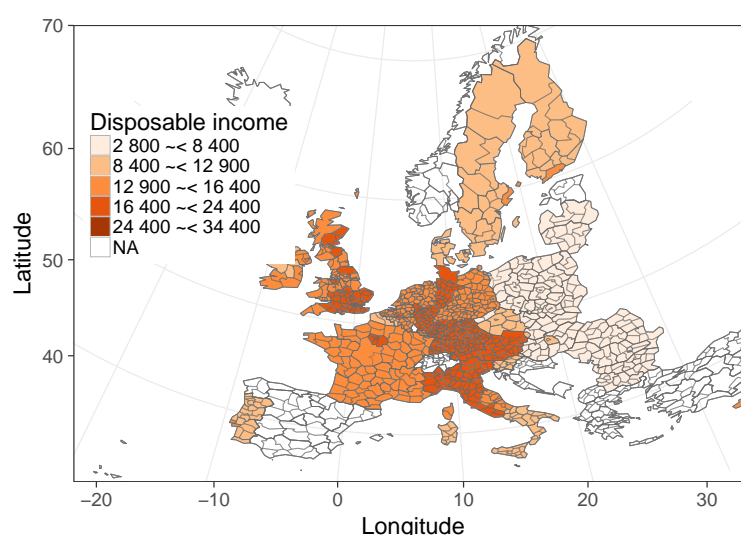


Figure 3: Disposable income of private households across NUTS2-level national regions in European countries. The household income statistics provided by Eurostat and the administrative boundaries by EuroGeographics© were obtained via the Eurostat open data service with the [eurostat](#) R package.

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

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

Similar lists are available for Euro area (`ea_countries`), EU (`eu_countries`) and the EU candidate countries (`eu_candidate_countries`). These auxiliary data sets facilitate smooth selection of specific country groups for a closer analysis. The full name and a two-letter identifier are provided for each country according to 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](#) package (Arel-Bundock, 2014).

Discussion

By combining programmatic access to data with custom analysis and visualization tools it is possible to facilitate a seamless automation of the complete analytical workflow from raw data to statistical summaries and final publication. The package supports automated and transparent data retrieval from institutional data repositories, featuring options such as search, subsetting and cache. Moreover, it provides several custom functions to facilitate the Eurostat data analysis and visualization. These tools can be used by researchers and statisticians in academia, government, and industry, and their applicability has been demonstrated in recent, independent publications (Kenett and Shmueli, 2016).

The [eurostat](#) R package provides a convenient set of tools to access open data from Eurostat, together with a comprehensive documentation and open source code via the package homepage. The documentation includes simple examples for individual functions, a generic package tutorial, and more advanced case studies on data processing and visualization. The package follows best practices in open source software development, taking advantage of version control, automated unit tests, continuous integration, and collaborative development (Perez-Riverol et al., 2016).

The source code can be freely used, modified and distributed under a modified BSD-2-clause license⁷. We value feedback from the user community, and the package has already benefited greatly from the user bug reports and feature requests, which can be systematically provided through the

⁷<https://opensource.org/licenses/BSD-2-Clause>

Github issue tracker⁸; advanced users can also implement and contribute new features by making pull requests. Indeed, these collaborative features have been actively used during the package development. We are committed to active maintenance and development of the package, and hope that this will encourage further feedback and contributions from the user community.

Acknowledgements

We are grateful to all package contributors, including François Briatte, Joona Lehtomäki, Oliver Reiter, and Wietse Dol, and to Eurostat for maintaining the open data service. This work is in no way officially related to or endorsed by Eurostat. The work has been partially funded by Academy of Finland (decisions 295741, 307127), and is part of rOpenGov⁹.

Bibliography

- V. Arel-Bundock. *WDI: World Development Indicators (World Bank)*, 2013. URL <http://CRAN.R-project.org/package=WDI>. R package version 2.4. [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. [p5]
- 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. *classInt: choose univariate class intervals*, 2015. URL <https://CRAN.R-project.org/package=classInt>. R package version 0.1-23. [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]
- E. Blondel. *rdsdmx: tools for reading SDMX data and metadata*, 2017. URL <https://CRAN.R-project.org/package=rdsdmx>. R package version 0.5-8. [p1]
- 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. [p1]
- M. J. A. Eugster and T. Schlesinger. OpenStreetMap and R. *R Journal*, 5(1):53–63, June 2012. [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]
- M. C. J. Kao, M. Gesmann, and F. Gheri. *FAOSTAT: download data from the FAOSTAT database of the food and agricultural organization (FAO) of the United Nations*, 2015. URL <http://CRAN.R-project.org/package=FAOSTAT>. R package version 2.0. [p1]
- D. R. S. Kenett and G. Shmueli. *Information quality: the potential of data and analytics to generate knowledge*. John Wiley & Sons, 2016. [p1, 5]
- L. Lahti, J. Parkkinen, and J. Lehtomäki. *statfi* R package, 2013a. [p1]
- L. Lahti, J. Parkkinen, J. Lehtomäki, and M. Kainu. rOpenGov: open source ecosystem for computational social sciences and digital humanities. Int'l Conf. on Machine Learning - ICML/MLOSS Open Source Software workshop, December 2013b. URL <http://ropengov.github.io>. [p1]
- J. Lemon. Plotrix: a package in the red light district of R. *R-News*, 6(4):8–12, 2006. [p3]
- M. Magnusson, L. Lahti, and L. Hansson. *pxweb: R tools for the px-web API*, 2014. URL <http://CRAN.R-project.org/package=pxweb>. R package version 0.5.57. [p1]

⁸<https://github.com/rOpenGov/eurostat/issues>

⁹<https://github.com/ropengov.io>

- R. McTaggart, G. Daroczi, and C. Leung. *Quandl: API wrapper for quandl.com*, 2015. URL <http://CRAN.R-project.org/package=Quandl>. R package version 2.7.0. [p1]
- J. Ooms. The jsonlite package: a practical and consistent mapping between JSON data and R objects. *arXiv:1403.2805 [stat.CO]*, 2014. URL <http://arxiv.org/abs/1403.2805>. [p1]
- E. Pebesma and R. Bivand. Classes and methods for spatial data in R. *R News*, 5(2), 2005. [p1]
- Y. Perez-Riverol, L. Gatto, R. Wang, T. Sachsenberg, J. Uszkoreit, F. da Veiga Leprevost, C. Fufezan, T. Ternent, S. J. Eglen, D. S. Katz, T. J. Pollard, A. Konovalov, R. M. Flight, K. Blin, and J. A. Vizcaíno. Ten simple rules for taking advantage of Git and GitHub. *PLoS Computational Biology*, 12(7):e1004947, July 2016. doi: 10.1371/journal.pcbi.1004947. [p5]
- 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]
- A. Reinhart. *pdfetch: fetch economic and financial time series data from public sources*, 2015. URL <http://CRAN.R-project.org/package=pdfetch>. R package version 0.1.7. [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. 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. *httr: tools for working with URLs and HTTP*, 2016. URL <https://CRAN.R-project.org/package=httr>. R package version 1.2.1. [p1]
- H. Wickham and R. Francois. *readr: read tabular data*, 2015. URL <https://CRAN.R-project.org/package=readr>. R package version 0.2.2. [p1]
- H. Wickham, R. Francois, and K. Müller. *tibble: simple data frames*, 2016. URL <https://CRAN.R-project.org/package=tibble>. R package version 1.1. [p1]

Leo Lahti
Department of Mathematics and Statistics
PO Box 20014 University of Turku
Finland
leo.lahti@iki.fi

Janne Huovari
Pellervo Economic Research PTT
Eerikinkatu 28 A, 00180 Helsinki
Finland
janne.huovari@ptt.fi

Markus Kainu
Research Department, The Social Insurance Institution of Finland
PO Box 450, 00101 Helsinki
Finland
markus.kainu@kela.fi

Przemysław Biecek
Faculty of Mathematics and Information Science
Warsaw University of Technology
Koszykowa 75, 00-662 Warsaw
Poland
P.Biecek@mimuw.edu.pl

Appendix

The full source code for this manuscript is available at the package homepage¹⁰. Source code for the obesity example (Figure 2A) is as follows.

```
library(dplyr)
tmp1 <- get_eurostat("hlth_ehis_de1", time_format = "raw")
tmp1 %>%
  dplyr::filter(isced97 == "TOTAL" ,
    sex != "T", age != "TOTAL", geo == "PL") %>%
  mutate(BMI = factor(bmi,
    levels=c("LT18P5", "18P5-25", "25-30", "GE30"),
    labels=c("<18.5", "18.5-25", "25-30", ">30"))) %>%
  arrange(BMI) %>%

  ggplot(aes(y = values, x = age, fill = BMI)) + geom_bar(stat = "identity") +
  facet_wrap(~sex) + coord_flip() +
  theme(legend.position = "top") +
  ggtitle("Body mass index (BMI) by sex and age") +
  xlab("% of population") + scale_fill_brewer(type = "div")
```

Source code for the renewable energy example (Figure 2B).

```
# All sources of renewable energy are to be grouped into three sets
dict <- c("Solid biofuels (excluding charcoal)" = "Biofuels",
  "Biogasoline" = "Biofuels",
  "Other liquid biofuels" = "Biofuels",
  "Biodiesels" = "Biofuels",
  "Biogas" = "Biofuels",
  "Hydro power" = "Hydro power",
  "Tide, Wave and Ocean" = "Hydro power",
  "Solar thermal" = "Wind, solar, waste and Other",
  "Geothermal Energy" = "Wind, solar, waste and Other",
  "Solar photovoltaic" = "Wind, solar, waste and Other",
  "Municipal waste (renewable)" = "Wind, solar, waste and Other",
  "Wind power" = "Wind, solar, waste and Other",
  "Bio jet kerosene" = "Wind, solar, waste and Other")

# Some cleaning of the data is required
energy3 <- get_eurostat("ten00081") %>%
  label_eurostat(dat) %>%
  filter(time == "2013-01-01",
    product != "Renewable energies") %>%
  mutate(nproduct = dict[as.character(product)], # just three categories
    geo = gsub(geo, pattern=" \\(.*", replacement="")) %>%
  select(nproduct, geo, values) %>%
  group_by(nproduct, geo) %>%
  summarise(svalue = sum(values)) %>%
  group_by(geo) %>%
  mutate(tvalue = sum(svalue), svalue = svalue/sum(svalue)) %>%
  filter(tvalue > 1000) %>%
  spread(nproduct, svalue)

# Triangle plot
positions <- plotrix::triax.plot(as.matrix(energy3[, c(3,5,4)]),
  show.grid = TRUE, label.points = FALSE, point.labels = energy3$geo,
  col.axis = "gray50", col.grid = "gray90",
  pch = 19, cex.axis = 1.1, cex.ticks = 0.7, col = "grey50")

ind <- which(energy3$geo %in% c("Norway", "Iceland", "Denmark", "Estonia", "Turkey", "Italy", "Finland"))
df <- data.frame(positions$xypos, geo = energy3$geo)
points(df$x[ind], df$y[ind], cex = 2, col = "red", pch = 19)
text(df$x[ind], df$y[ind], df$geo[ind], adj = c(0.5, -1), cex = 1.5)
```

¹⁰<http://ropengov.github.io/eurostat>