Tempo – an R package to access the TEMPO-Online database

Marian Necula (neculamarian18@stud.ase.ro)
The Bucharest University of Economic Studies,

Ana Maria Țîru (ana.tiru@insse.ro)
National Statistics Institute

Bogdan Oancea (bogdan.oancea@faa.unibuc.ro)
University of Bucharest

ABSTRACT

One of the objectives of official statistical authorities is to extend and diversify the methods used to disseminate data, in order to satisfy various and ever-changing user needs. This paper presents the second iteration of Tempo, an R package that provides access to statistical information stored under Romanian National Institute of Statistics (NSI) public database, [1]TEMPO-Online. Tempo provides a collection of custom functions to download, explore and visualize statistical data in an automated and convenient manner. Access to data is enabled by TEMPO-Online Service Application Programming Interface (API), which exposes JSON-based datasets and metadata. To access the API from R, Tempo relies on "curl" package. The Tempo package also brings in a visualization feature based on "ggplot2" package. This feature allows the user to render a selected dataset as a gradient mesh over the Romanian country map.

Keywords: *R, Tempo Online, open data, data dissemination*
JEL Codes: *C88, C89*

1. INTRODUCTION

Open data is data generated or collected by public authorities, "data that can be freely used, reused and redistributed by anyone" (General Secretariat of the Government, 2019). The European Commission's Open Data activity focuses on generating value through the use and re-use of public sector information (government data), encouraging innovation and successful economic initiatives.

Open data meets the following features: it is readily and widely accessible, can be reused, it is available online in a file format that can be

automatically processed using computer programs (machine-readable) and it is available to anyone in bulk and free of charge. The NSI is part of the initiative to open public data through information dissemination activities, e.g. developing and publishing the R package Tempo. This package offers access to open data by bulk downloading statistics, free of charge, using an automated process. The downloaded information is user friendly, easy to understand and has a standard format that allows easy integration into other information or analysis systems, and can be modified and reused.

R environment is a suite of software functionalities built for complex data manipulation, mathematical and statistical operations, and graphical display. It is used mainly for research in various fields like biology, medicine, economics, agriculture or business. The set of functionalities provided by R can be extended using R packages. Any user can create the package and has the option to share it to a public repository. Within an R environment, a set of packages are already enabled by default, including "base", "compiler", "graphics", "stats". A public repository like CRAN ("Comprehensive R Archive Network") can be used to download, install and enable a vast list of packages like "ggplot2", "curl", "knitr" etc.

Some European statistical offices already have developed R packages useful for accessing official statistics. For example, the R package "cbsodataR" (De Jonge, 2016) allows users to access the open data of Statistics Netherlands, the R package "sorvi" (Lahti, 2016) offers tools to download and manipulate Finnish Open Government Data, the R package "eurostat" (Lahti, 2019; Lahti et al. 2017) brings in tools to download data from the Eurostat database together with search and manipulation utilities. Other R packages developed by non-European producers of official statistics such as Statistics Canada, U.S. Census Bureau, Mexico's Official Statistics Agency that facilitate access to the open statistical data are: "CANSIM2R" (Lugo, 2018), "censusapi" (Recht, 2018), "acs" (Glenn, 2019), "inegiR" (Flores, 2018)

The process of statistical analysis requires access to data sources. At country level, the Romanian National Institute of Statistics manages one of the most complex database of statistical data, called TEMPO-Online, made available to the public via TEMPO-Online web service. The content of TEMPO-Online database comprises of statistical indicators, metadata associated with statistical indicators (definition, methodology) and continuous time series that begin with 1990 (with monthly, quarterly and annual periodicity). Outside of the R environment, users can access this data source using a web browser to search and download datasets and can opt to export the data tables as Excel or .csv files. This process can be automatized by using R and this is the main purpose of the Tempo R package.

2. METHODOLOGY

R packages are collections of R functions, data sets, documentation and tests, which provide a transparent way of sharing code, and are mainly distributed through CRAN or Github Web sites as open source projects. The package system allows outside contribution and development to R environment while still imposing some basic standards. Another positive feature of R Packages is given by the possibility to be dynamically loaded and unloaded at runtime, the memory being occupied only when the package is actually used.

Tempo package was developed using the R packaging system and it relies on the API enabled by TEMPO-Online Service of NSI. The package aims to help users by offering a facile access to TEMPO-Online database by bulk downloads in .csv format.

Compared to the previous version of rTempo package, a different technique for accessing the data is used, namely through the API provided by TEMPO-Online Service. Using API ensures consistency of data download processes that is essential for recurring statistical analyses.

Tempo package was built on two packages: *curl* (Ooms, 2019) and *jsonlite* (Ooms, 2018). Tempo also brings in a visualization feature, based on ggplot2 (Wickham, 2019; Wickham, 2016) package.

curl package is used to ensure the underlying connection between R and the Web. It contains functions that facilitate URL querying and data fetching from the Web, allowing to store the response data into memory or disk for later processing. Out of all *curl* package functions the following were used for developing the package: *curl_fetch_memory*, *handle* and *parse_headers*. The *curl* package provides rich configuration in order to build the request, like custom options, headers, or payload.

jsonlite package provides functions used for converting JSON data from/to R objects and interacting with a Web API. An example of high-level interaction with *curl* and *jsonlite* packages consists of querying a predefined URL using a *curl* function that builds the HTTP request (usually GET or POST), then getting a response from the server as a JSON string, and then transforming the response string into an R object (dataframe) using *jsonlite*.

The R Tempo package acts as a thin client over TEMPO-Online database, providing methods to query and retrieve statistical data ready to use within an R environment. To query the Tempo-Online database, Tempo relies on a Web service called TEMPO-Online Service, available over HTTP at http://statistici.insse.ro:8077/tempo-ins. The Web service exposes a REST API with methods that allow GET or POST requests, and return structured

data in JSON format. The API methods are divided in 3 categories: statistical domains (available at "/context"), dataset metadata (available at "/matrix"), and dataset content (available at "/matrix/dataSet").

Data visualization is a vital mechanism used to gain meaningful insights from data. *ggplot2* authored by Hadley Wickham is a visualization package based on the grammar of graphics theory. *ggplot2* is an elegant and powerful tool used to create complex plots. This package builds graphics following a layered approach and combining a data set, a set of geometric objects (points, lines, polygons), aesthetics (color, shape, size) and a coordinates system. The data in TEMPO Online database is commonly available as annual time series at NUTS (Nomenclature of Territorial Units for Statistics) level by counties, regions and macroregions. *ggplot2* is used to visualize the data offered by TEMPO Online database at NUTS level on the Romanian country map.

3. TEMPO IMPLEMENTATION

TEMPO package provides an access point from R programming environment for downloading and manipulating TEMPO Online datasets. The package's backbone is the TEMPO Online Service Application Programming Interface (Romanian National Statistics Institute, 2018). The following TEMPO package functions are available for the R user:

set_language(language = c("ro", "en"))

This function allows the users to set the language used for downloading a table from TEMPO Online. The language parameter is a string to set the language for the downloaded tables. Options: "ro" - for Romanian and "en" - for English. If no parameter is given, implicitly downloads tables in Romanian.

> set_language(language = "en")

get_language()

This function allows the users to get the language used for downloading a table from TEMPO ONLINE.

> get_language()

[1] "en"

tempo_toc(fullDescription = FALSE)

This function downloads the *Table of Contents* (TOC) for Tempo Online database. If the parameter *fullDescription* is TRUE then the *tempo_toc* function starts to collect dates for last updates.

>tempo_toc(fullDescription = FALSE)

name	*code*
Colective accidents by economic activities CANE Rev.1	*ACC101A*
Colective accidents at work by macroregions, development regions and counties	*ACC101B*
Colective accidents at work by economic activities CANE Rev.2	*ACC101C*
Injuried persons at work by economic activities CANE Rev.1, type of accidents at work	*ACC102A*
Injuried persons at work by type of accidents at work, by macroregions,regions and counties	*ACC102B*
Injuried persons at work by economic activities CANE Rev.2, type of accidents at work	*ACC102C*

tempo_bulk(codes = NULL, directory = NULL, check_date = FALSE)

Download one or multiple tables from TEMPO Online database (bulk download). The function checks the directory supplied as parameter or the working directory if the files names are identical to TEMPO Online matrices. If the files from working directory are more recent than the last date upon which the TEMPO matrices have been updated, the function skips them at download.

Function parameters:

- *codes* - a list containing one or more strings containing the code for the table/matrix in TEMPO Online database.
- *directory* - a valid path where the files will be downloaded. The implicit value is NULL, i.e. current working directory.
- *check_date* - if FALSE (by default) the files are downloaded and overwritten, if TRUE the date of the files from the directory are compared to the latest update of TEMPO matrices and only the files downloaded before the latest update of TEMPO are downloaded.

tempo_filter()

This function returns a list of vector strings containing values for subsetting a Tempo Online matrix.

> tempo_filter("AGR111A")

tempo_options()

This function returns a list of named integers representing the codes for subsetting a Tempo Online matrix.

> tempo_options("AGR111A")

tempo_clean(matrix)

tempo_clean function allows users to clean a table downloaded from TEMPO Online database through *tempo_bulk* function, by removing redundant columns.

```
>head(AMG157G)
   Age.group      Sex      Months          MU..Percentage    Value
1  15 - 24 years  Total    January 2004    Percentage        20.2
2  15 - 24 years  Total    February 2004   Percentage        20.2
3  15 - 24 years  Total    March 2004      Percentage        20.2
4  15 - 24 years  Total    April 2004      Percentage        21.1
5  15 - 24 years  Total    May 2004        Percentage        21.1
6  15 - 24 years  Total    June 2004       Percentage        21.1
>tempo_clean(matrix = AMG157G)
      Age.group       Sex      Months          Value/ Percentage
1     15 - 24 years   Total    January 2004    20.2
2     15 - 24 years   Total    February 2004   20.2
3     15 - 24 years   Total    March 2004      20.2
4     15 - 24 years   Total    April 2004      21.1
5     15 - 24 years   Total    May 2004        21.1
6     15 - 24 years   Total    June 2004       21.1
…
```

tempo_search(keyword = c())

tempo_search function downloads the *"Table Of Contents"* (TOC) for Tempo Online database and returns a TOC subset based on specified keywords. An example of a call for *tempo_search* function is given below:

```
>tempo_search(c("education", "industry"))
```

tempotime2date(matrix)

tempotime2date function gives the possibility to convert a column of class character representing time periods (month, quarter) from a table downloaded through *tempo_bulk* function to a column of class date representing calendar dates. If the downloaded data does not contain time values, the function returns the original data. The parameter matrix is an R object with time information in TEMPO time format, representing the table/ matrix downloaded from TEMPO Online database. For example, a call to *tempotime2date()* will look like this:

```
>head(AMG157G)
     Age.group   Sex       Months      MU..Percentage Value
1 15 - 24 years  Total  January 2004    Percentage    20.2
2 15 - 24 years  Total  February 2004   Percentage    20.2
3 15 - 24 years  Total   March 2004     Percentage    20.2
4 15 - 24 years  Total   April 2004      Percentage    21.1
5 15 - 24 years  Total    May 2004       Percentage    21.1
6 15 - 24 years  Total    June 2004      Percentage    21.1
>tempotime2date(matrix = AMG157G)
     Age.group   Sex      Months     MU..Percentage Value
1 15 - 24 years  Total  2004-01-01    Percentage    20.2
2 15 - 24 years  Total  2004-02-01    Percentage    20.2
3 15 - 24 years  Total  2004-03-01    Percentage    20.2
4 15 - 24 years  Total  2004-04-01    Percentage    21.1
5 15 - 24 years  Total  2004-05-01    Percentage    21.1
6 15 - 24 years  Total  2004-06-01    Percentage    21.1
…
```

tempo_geo(matrix,year,area,filter,title)

tempo_geo function is used to plot the data downloaded through *tempo_bulk* function on the Romanian country map. If the downloaded data are not grouped in counties, regions or macroregions, then *tempo_geo* returns an error. The user must specify a filter for each variable factor present in the data set. The parameters *matrix*, *year*, *area* and *filter* are mandatory to be identified and set up from the beginning. The parameter *title* is optional. If the parameter *title* is not specified, the function will set a default title based on the filter and year selected by the user. If there is no data for the year or the area specified by the user, then *tempo_geo* will return an error and the existing values from which the user can choose. The result obtained, in the form of a map, is saved on disk in .png format. For a better understanding of the parameters selection, an example is given here below:

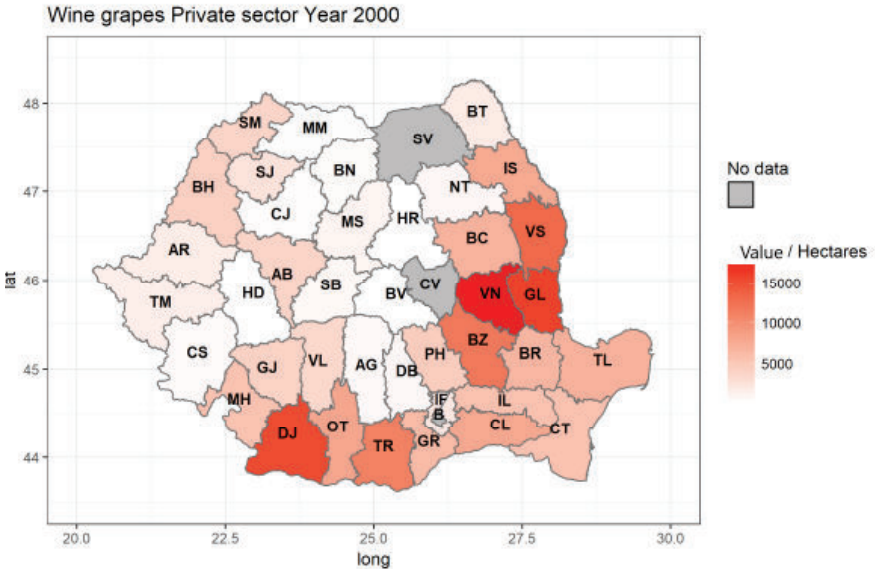
- matrix: the table/matrix downloaded from TEMPO Online database containing geospatial data;
- year: a string containing the year used to plot geospatial data;
- area: a string to set the type of area (counties, regions, macroregions) used to group geospatial data;
- filter: character vector containing values from factor variables;
- title: a string to set the map title.

The following call of the *tempo_geo* function will return the maps presented in Figure 1, Figure 2 and Figure 3.

```
>tempo_geo(matrix = AGR111A,
      year = "2000",
      area = "counties",
      filter = c("Wine grapes","Private sector"))
>tempo_geo(matrix = AGR111A,
      year = "2000",
      area = "regions",
      filter = c("Wine grapes","Private sector"))
>tempo_geo(matrix = AGR111A,
      year = "2000",
      area = "macroregions",
      filter = c("Wine grapes","Private sector"))
```

Romanian country map devided by counties

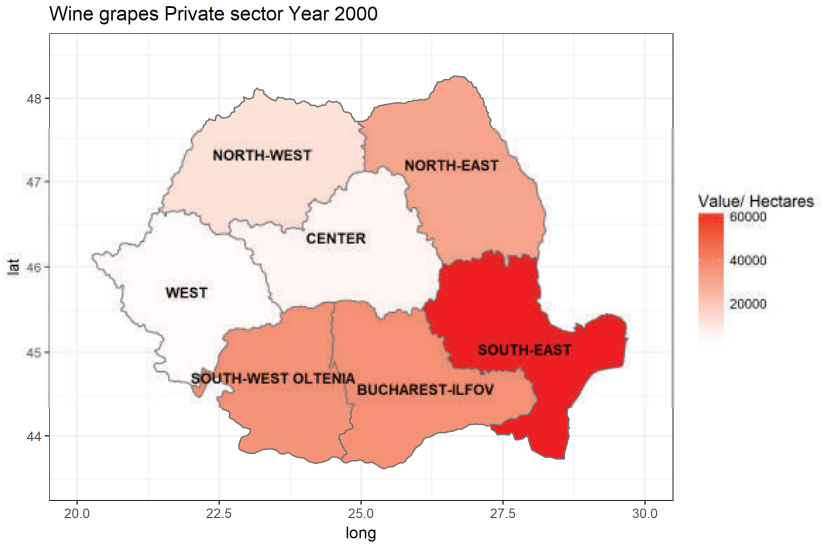
Figure 1



Source: designed by the authors

Romanian country map divided by regions

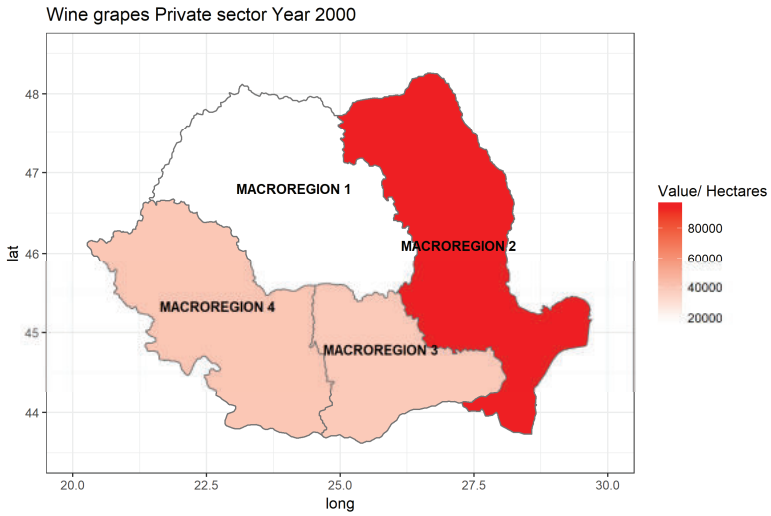
Figure 2



Source: designed by the authors

Romanian country map divided by macroregions

Figure 3



Source: designed by the authors

CONCLUSIONS

The Tempo package is highly flexible offering a convenient set of custom functions to access open data from TEMPO online database. Researchers and data scientists in academia, government, and industry can use these functions to download, explore and visualize statistical data in an automated and convenient manner. Foreseen usage of the package can be categorized as: automated data pipelines for analytics, data visualization and exploration, or data download as an independent step for future processing.

Tempo package takes the Romanian Institute of Statistics one-step further in the European endeavor to provide modern tools and a cohesive environment for statistics while working with open data.

REFERENCES

1. **Dușa, A., Oancea, B., Caragea N., Alexandru, C., Jula, N.M., Dobre, A.M.,** 2015, *R cu aplicații în statistică*. Bucharest: The Bucharest University Press.
2. **Eurostat**, 2011, *European Statistics Code of Practice*, Luxembourg: European Statistical System.
3. **Flores**, **E.**, 2018, *inegiR*, R package version 2.0.0 [Online] Available at: *https://cran.r-project.org/package=inegiR.*
4. **Glenn**, **E. H.**, 2019, *acs - Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census*, R package version 2.1.4, [Online] Available at: *https://cran.r-project.org/package=acs*
5. **De Jonge, E**., (2016), *cbsodataR: Statistics Netherlands (CBS) Open Data API Client*, version 0.2.1, [Online] Available at: https://cran.r-project.org/package=cbsodataR
6. **Leisch, F.,** 2009, *Creating R Packages: A Tutorial,* in *Brito, P.* (Editor), Compstat 2008 – Proceedings in Computational Statistics. Physica Verlag, Heidelberg.
7. **Lahti, L.**, 2019, *eurostat: Tools for Eurostat Open Data.* R package version 3.3.5, [Online] Available at: *https://cran.r-project.org/web/packages/eurostat/eurostat.pdf* [Accessed 6 May 2019].
8. **Lahti, L., Huovari, J., Kainu, M., Biecek, P.**, 2017, *"Retrieval and analysis of Eurostat open data with the eurostat package"*, R Journal 9(1):385-392.
9. **Lahti, L., Parkkinen, J., Lehtomaki, J., Haapanen, J., Happonen, E., and Paananen, J.,** 2015. *sorvi: Finnish open data toolkit for R*, [Online] Available at: *http://ropengov.github.com/sorvi*
10. **Lugo, M.,** 2018, *CANSIM2R - Directly Extracts Complete CANSIM Data Tables*. R package version 1.14.1 [Online] Available at: *https://cran.r-project.org/web/packages/CANSIM2R/index.html*
11. **Ooms, J.**, 2019, *curl: A Modern and Flexible Web Client for R.* R package version 3.3.0, [Online] Available at: *https://cran.r-project.org/web/packages/curl/index.html*
12. **Ooms, J.**, 2018, *jsonlite: A Robust, High Performance JSON Parser and Generator for R.* R package version 1.6 [Online] Available at: *https://cran.r-project.org/web/packages/jsonlite/index.html*
13. **R Core Team**, 2017, *Writing R extensions. Pages: 3-17*. [Online] Available at: *https://cran.r-project.org/doc/manuals/R-exts.pdf.*
14. **Recht, H**., 2018, *censusapi*, R package version 0.6.0 [Online] Available at: *https://cran.r-project.org/package=censusapi*

15. **Secretariatul General al Guvernului,** 2018, *Metodologie pentru publicarea datelor deschise. Creșterea calității și a numărului de seturi de date deschise publicate de instituțiile publice - cod SIPOCA 36*. [Online] Available at: *http://ogp.gov.ro/wp-content/uploads/2018/06/Metodologie-date-deschise_iunie2018.pdf*
16. **Wickham, H.,** 2019, *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* R package version 3.1.1 [Online] Available at: *https://cran.r-project.org /web/ packages/ggplot2/index.html*
17. **Wickham**, **H.**, 2016, *"ggplot2: Elegant Graphics for Data Analysis"*, Springer-Verlag New York.