2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8345

Access and analysis of ISTAC data through the use of R and Shiny

González-Martel, Christian[a] ; Cazorla Artiles, José M.[b]; Pérez-González, Carlos J.[c]
[a]Departamento de Métodos Cuantitativos en Economía y Gestión, Universidad de Las Palmas de Gran Canaria, Spain, [b]Universidad de Las Palmas de Gran Canaria, Spain, [c]Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Spain.

Abstract

The increasing availability of open data resources provides opportunities for research and data science. It is necessary to develope tools that take advantage of the full potential of new information resources. In this work we developed the package for R istacr that provides a collection of eurostat functions to be able to consult and discard the data that Eurostat, including functions to retrieve, download and manipulate the data set available through the ISTAC BASE API of the Canary Institute of Statistics (ISTAC). In addition, A Shiny app was designed for a responsive visulization of the data. This develope is part of the growing demand for open data and ecosystems dedicated to reproducible research in computational social science and digital humanities. With this interest, this package has been included within rOpenSpain, a project that aims to promote transparent research methods mainly through the use of free software and open data in Spain.

Keywords: *Economic databases; R; package; Shiny; visualization.*

1. Introduction

The Open Data initiative or data opening is a practice that seeks to ensure that certain data and information belonging to public administrations and organizations are accessible and available to everyone, without technical or legal restrictions.

Ruijer et al. (2017) have studied a context-sensitive open data design that facilitates the transformation of raw data into meaningful information constructed collectively by public administrators and citizens. Thorsby et al. (2017) research on features and content of open data portals in American cities. Their results show that, in general, the portals are in a stage of development and need to improve user help and analysis features as well as inclusion of features to help citizens understand the data, such as more charting and analysis.

The reproducible research defined as the complete analytical workflows, fully replicable and transparent, that span from raw data to final publications can benefit from the availability of algorithmic tools to access and analyse open data collections (Gandrud, 2013; Boettiger et al., 2015). Dimou et al. (2014) presents a use case of publishing research metadata as linked open data and creating interactive visualizations to support users in analyzing data in a research context.

However, the data provided in open access are not in a standardized format and arises the need to adapt the code to specific data sources to accommodate variations in raw data formats, access details so that the end users can avoid repetitive programming tasks and save time allowing simplification, standardization, and automation of analysis workflows facilitating reproducibility, code sharing, and efficient data analytics.

Following this idea, within the ecosystem of R, several packages have been created to work with data from Food and Agricultural Organization (FAO) of the United Nations (FAOSTAT; Kao et al. 2015), World Bank (WDI; Arel-Bundock 2013, wbstats; Jesse Piburn 2018), Open Street Map (osmar; Eugster and Schlesinger 2012) amog others.

The Canary Institute of Statistics (Instituto Canario de Estadística, ISTAC) provides a rich collection of data, including thousands of data sets on Canarian demography, health, employment and tourism and other topics in an open data format.

ISTAC is the central authority of the canary statistical system and the official research office of the Government of the Canary Islands and, among others, among its functions are to provide statistical information and coordinate the public statistical activity of Canary Island autonomous region.

The main access to ISTAC is the web-based graphical user interface (GUI) from where the data can be consulted and downloaded in alternative formats. This access method is fine for the occasional use but is tedious for large selections and when the user must access to data very frequently. The second method uses an Application Programming Interface (API) that can be embedded in a computer code to programmatically extract data from ISTAC. We have developed a R package that integrates the API into the code that allows for the downloaded data to be directly manipulated in R. Based on this package, we have also created a Shiny application that allows a visualization of ISTAC data.

The visualization characteristics is one of the most important features in analyzing information from open data sources. Chen and Jin (2017) have recently proposed a data model and application procedure that can be applied for visualization evaluation and data analysis in human factors and ergonomics. Jones et al. (2016) research innovative data visualization and sharing mechanisms in the study of social science survey data on environmental issues in order to allow the participatory deliberation. Kao et al. (2017) shows how to use a visualization analysis tool for open data with the aim to verify whether there exists sensitive information leakage problem in the target datasets.

This paper provides an overview of the core functionality in the current release version. A comprehensive documentation and source code are available via the package homepage in Github[1]. The package is part of rOpenSpain[2], an initiative whose objective is to create R packages to exploit open data available in Spain for reproducible research.

This paper is structured as follows: firstly, we explain the data extraction procedure implemented in the R library and the workflow to achieve visualization of data. In section 3 we explain the architecture of the visualizations with Shiny. Finally, we present some concluding remarks.

2. The extraction routine in istcar

To install and load the last release version of istacr, the user should type in R the installation from GitHub command from the devtools package.

```
devtools::install_github(“rOpenSpain/istacr”)
```

```
library(“istacr”)
```

When the package is loaded the metadata of each dataset available by ISTAC BASE API are also loaded into the `cache` variable. It contains information about the title, topic, subtopic, the url to access to the json data, among other.

For searching about a specific term the `istac_search()` function is provided.

[1] https://github.com/rOpenSpain/istacr

[2] https://ropenspain.es/

```
busqueda.egt <- istac_search("egt", fields = "datos publicadosII")
```

This seeks among all the ISTAC BASE datasets those in which the pattern “egt” appears within the field “datos publicados II”. Other fields can be “titulo” (default), “tema”, “subtemaI”, “subtemaII”, “datos publicados I”, “origen” and ”encuesta”. You can obtain the list of fields with

```
names(cache)
```

The patter can be used with regular expression operators. The output are the rows or row of cache that keep to the pattern. Values in the ID column of the output provide data identifiers for subsequent download commands.

```
busqueda.egt$ID[1]
## [1] "sec.hos.enc.ser.2528"
```

2.1. Downloading data from ISTAC

We retrieve the data from the dataset with the ID reference using the ISTAC BASE API.

```
df <- istac(busqueda.egt$ID[1])
```

By default the function istacr works with human-readable labels. With the argument `label = FALSE` the function converts the labels into less interpretable codes.

The indicators in the ISTAC open data service are typically available as annual time series grouped by islands, but sometimes at a different granularity or geographic levels.

If the dataset has the “Islas” column it can be filtered by islands using the argument `islas = TRUE`, otherwise this argument is ignored. Valid values for islands are: El Hierro, La Palma, La Gomera, Tenerife, Gran Canaria, Fuerteventura and Lanzarote.

The function allows filtering the dataset by dates using the arguments `startdate`, `enddate`, and `mrv`. The argument `freq` controls the granularity of the data for fetching yearly ("anual"), biannual ("semestral"), quaterly ("trimestral"), monthly("mensual"), bi-weekly("quincenal"), weekly("semanal") values.

2.2. Data visualization

Istacr by itself does not have a dedicated function to plot the data but you can use the potential that R provides to visualize the data retrieved. Figure 1 shows the result of the combination of istacr and ggplot2 package.

```
ggplot (df_p, aes(x = Periodos, y = valor, fill = `Países de
residencia`)) +
      geom_col() +
      facet_wrap(~`Países de residencia`) +
      theme_bw()+
      theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
labs(x = "Years", y = "Total expenditure (euros)", title = "Total
tourist expenditure according to countries of residence")
```

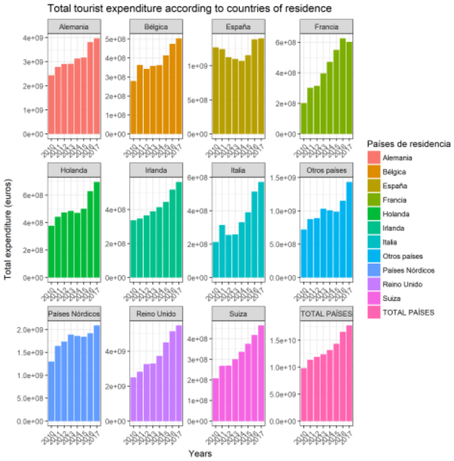


Figure 1. Data obtained through istacr. Visualization with ggplot function from ggplot2 package. Source:ISTAC (2018).

Because that most of ISTAC dataset contains geographical information, map visualization can be represented in a very natural way.

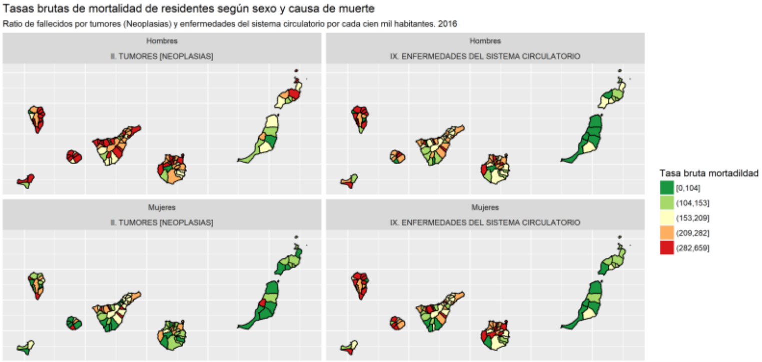

Tasas brutas de mortalidad de residentes según sexo y causa de muerte
Ratio de fallecidos por tumores (Neoplasias) y enfermedades del sistema circulatorio por cada cien mil habitantes. 2016
Hombres
II. TUMORES [NEOPLASIAS]
Hombres
IX. ENFERMEDADES DEL SISTEMA CIRCULATORIO
Mujeres
II. TUMORES [NEOPLASIAS]
Mujeres
IX. ENFERMEDADES DEL SISTEMA CIRCULATORIO
Tasa bruta mortadildad
[0,104]
(104,153]
(153,209]
(209,282]
(282,659]

Figure 2. Geographical visualization with ggplot2 package. Source: Análisis de la Mortalidad/ Series anuales. Municipios de Canarias. 1999-2016. ISTAC.

3. Creating a Shiny Web Application

The last step in this work focuses in generate a web application in R (Shiny). The Shiny is feed up using the ISTAC base API through istacr and its main purposes are the access to ISTAC tourism data and to facilitate the understanding of tourism patterns in the Canary Islands with special attention to exploit the statistics which comes from Tourism Expenditure Survey.

Based on the idea from New Zealand Tourism Dashboard[3], the Canary Islands Tourism Dashboard has been developed[4,5]

The structure of this Shiny shows a navigation menu upper bound. This menu contains several sections. The first one is a brief description of the application. The second one is the tourism expenditure section, this section is also composed subsections related and showed by a dropdown menu. The third sections is referred to the tourist profile socio-demographics characteristics are shown here. The fourth section is about the travel characteristics. Last section is about the authors. In future the purpose it is to improve this options to get a more complex application.

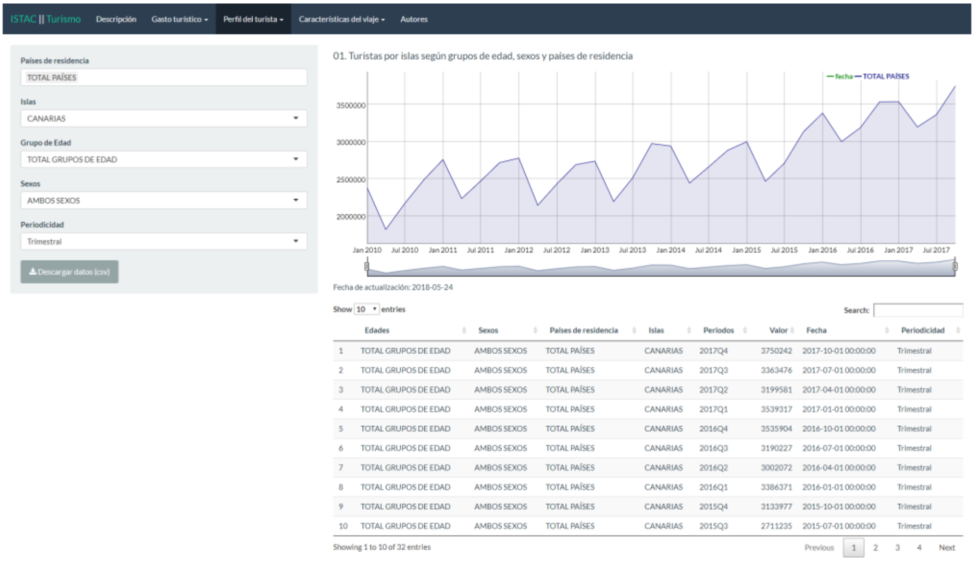
An important reason to use Shiny is the interaction between user and server. In this sense the user can filter data and change the rendered visualization. The Shiny application has an option to download data and export visualization are also available.

[3] https://mbienz.shinyapps.io/tourism_dashboard_prod/

[4] https://jmcartiles.shinyapps.io/canary_islands_tourism_dashboard/

[5] Full code aviable on https://github.com/jmcartiles/canary_islands_tourism_dashboard

A Shiny example of use is shown in Figure 3. The data represented is referred to the number of tourist quarterly to the Canary Islands by age, sex and residence country. In this tab, the user has four filter options in a dropdown menu at left-side. The results are displayed as chart and table. In the data panel a search option to filter data by pattern and a sorting option could be also used.



Países de residencia: TOTAL PAÍSES

Islas: CANARIAS

Grupo de Edad: TOTAL GRUPOS DE EDAD

Sexos: AMBOS SEXOS

Periodicidad: Trimestral

Fecha de actualización: 2018-05-24

Show 10 entries

Search:

	Edades	Sexos	Países de residencia	Islas	Periodos	Valor	Fecha	Periodicidad
1	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q4	3750242	2017-10-01 00:00:00	Trimestral
2	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q3	3363476	2017-07-01 00:00:00	Trimestral
3	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q2	3199581	2017-04-01 00:00:00	Trimestral
4	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q1	3539317	2017-01-01 00:00:00	Trimestral
5	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q4	3535904	2016-10-01 00:00:00	Trimestral
6	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q3	3190227	2016-07-01 00:00:00	Trimestral
7	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q2	3002072	2016-04-01 00:00:00	Trimestral
8	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q1	3386371	2016-01-01 00:00:00	Trimestral
9	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2015Q4	3133977	2015-10-01 00:00:00	Trimestral
10	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2015Q3	2711235	2015-07-01 00:00:00	Trimestral

Showing 1 to 10 of 32 entries

Figure 3. Shiny application for istacr data visualization. Source:ISTAC (2018).

4. Conclusions

In the last years have emerged a high numer of tools that enable the sharing of public data in open formats across different cloud platforms through API web services. One of the problems that arises is how to collect data of interest from such sources. The istacr library allows the users to query and obtain statistical data series in an efficient and convenient way. The main aspect of this library consists in connecting to API web service of ISTAC to access data and, then, create a dataset into R to work with it. This represents an advantage respect to other procedures that download data in a computer file previously to read from R.

In most of cases, the visualizations are used to demonstrate the provided information in a alternative fashion to the information they present. The visualizations can provide some significant insights of the open data and allow to non-expert users the opportunities discovery in their data analyzes. Therefore, the usefulness of Open Data is revealed to non-expert users. In this use case, it is highlighted how Open Data helps in improving the quality of the data, the diversity of the information and the integration of knowledge.

Considering these visualizations, the potential that offers istacr package could be highly interesting for managing this kind of data.

References

Arel-Bundock, V. (2013) WDI: World Development Indicators (World Bank), URL http://CRAN.R-project.org/package=WDI

Boettiger, C., Chamberlain, S., Hart, E., Ram, K. (2015). Building software, building community: lessons from the rOpenSci project, *Journal of Open Research Software*, 3(1), DOI http://doi.org/10.5334/jors.bu

Chen, X. & Jin, R. (2017) Statistical modeling for visualization evaluation through data fusion, *Applied Ergonomics*, 65, 551-561.

Dimou,A., De Vocht, L., Van Grootel, G., Van Campe, L., Latour, J., Mannens, E., Van de Walle, R. (2014) Visualizing the Information of a Linked Open Data Enabled Research Information System, *Procedia Computer Science*, 33, 245-252

Eugster, M. J. A. and Schlesinger, T. (2010) osmar: OpenStreetMap and R, URL http://osmar.r-forge.r-project.org/RJpreprint.pdf

Gandrud, C. (2013). *Reproducible Research with R and RStudio.* Chapman & Hall/CRC

Jones, A. S., Horsburgh, J. S., Jackson-Smith, D., Ramírez, M., Flint, C. G., Caraballo, J. (2016) A web-based, interactive visualization tool for social environmental survey data, *Environmental Modelling & Software*, 84, 412-426.

Kao,C.-H., Hsieh, C.-H., Chu, Y.-F., Kuang, Y.-T., Yang, C.-K. (2017) Using data visualization technique to detect sensitive information re-identification problem of real open dataset, *Journal of Systems Architecture*, 80, 85-91.

Kao, M.C.J., Gesmann, M., Gheri, F. (2015). FAOSTAT: Download Data from the FAOSTAT Database of the Food and Agricultural Organization (FAO) of the United Nations, URL https://cran.r-project.org/web/packages/FAOSTAT/index.html

Piburn, J. (2018). wbstats: Programmatic Access to the World Bank API, URL https://www.ornl.gov/division/csed/gist

Ruijer, E., Grimmelikhuijsen, S., Meijer, A. (2017) Open data for democracy: Developing a theoretical framework for open data use, *Government Information Quarterly*, 34(1), 45-52.

Thorsby, J., Stowers, G.N.L., Wolslegel, K., Tumbuan, E. (2017) Understanding the content and features of open data portals in American cities, *Government Information Quarterly*, 34(1), 53-61.