covid19.analytics: An R Package to Obtain, Analyze and Visualize Data from the 2019 Corona Virus Disease Pandemic

21 January 2021

Summary

With the emergence of a new pandemic worldwide, a novel strategy to approach it has also emerged. Several initiatives under the umbrella of *open science* are contributing to tackle this unprecedented situation. In particular, the *R Language and Environment for Statistical Computing* (R Core Team 2016; Ihaka and Gentleman 1996) offers an excellent tool and ecosystem for approaches focusing on open science and reproducible research. Hence it is not surprising that with the onset of the pandemic, a large number of R packages and resources were made available for researches working in the pandemic.

In this paper, we present the `covid19.analytics` R package which allows users to access and analyze worldwide data from publicly available resources.

The package also provides basic analysis and visualization tools and functions to investigate these datasets and other ones structured in a similar fashion.

Statement of need

The `covid19.analytics` package is an open source tool, which its main implementation and API is the R package (Ponce 2020) which its stable version is part of CRAN (*The Comprehensive R Archive Network*, n.d.). In addition to this, the package has a few more adds-on:

- a central GitHUB repository, https://github.com/mponce0/covid19.analytics where the latest development version and source code of the package are available. Users can also submit tickets for bugs, suggestions or comments using the "issues" tab.
- a rendered version with live examples and documentation also hosted at GitHUB pages, https://mponce0.github.io/covid19.analytics/

- a dashboard for interactive usage of the package with extended capabilities for users without any coding expertise, https://covid19analytics.scinet.utoronto.ca
- a **backup** data repository hosted at GitHUB, https://github.com/mponce0/covid19analytics.datasets – where replicas of the live datasets are stored for redundancy and robust accessibility sake (see Figure 1).

The `covid19.analytics` R package allows users to obtain live[1] worldwide data of reported cases from the novel *Corona Virus Disease* (CoViD19), as well as related datasets, such as, genomics data, vaccinations records and historical pandemics records. It does this by accessing and retrieving the data publicly available and published by several sources:

- the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" ("COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" 2020) for the worldwide and US data
- Health Canada ("Health Canada – covid19 cases in Canada" 2020), for Canada specific data
- the city of Toronto for Toronto data ("COVID-19: Status of Cases in Toronto" 2020)
- Open Data Toronto for Toronto data ("Open Data Toronto – covid19 cases in the City of Toronto" 2020)
- Our World In Data for vaccination data (Our World In Data 2020)
- NCBI servers for genomics datasets (NCBI Resource Coordinators 2017)
- Visual Capitalist infographics for historical pandemic records (Visual Capitalist 2020)

The package also incorporates a dashboard to facilitate the access to its functionalities to less experienced users (see Figure 2).

As today, there are a few dozen other packages also in the CRAN repository that allow users to gain access to different datasets of the CoViD19 pandemic. In some cases, some packages just provide access to data from specific geographical locations or the approach to the data structure in which the data is presented is different from the one presented here. Nevertheless, having a variety of packages from which users can try and probably combine, is an important and crucial element in data analysis. Moreover it has been reported different cases of data misuse/misinterpretation due to different issues, such as, erroneous metadata or data formats (Ledford and Van Noorden 2020) and in some cases ending in articles' retractions (Schriml et al. 2020). Therefore providing additional functionalities to check the integrity and consistency of the data, as our `covid19.analytics` package does is paramount. This is specially true in a situation where the unfolding of events and data availability is flowing so fast that sometimes is even hard to keep track of all the changes.

[1]The data usually is accessible from the repositories with a 24 hours delay.

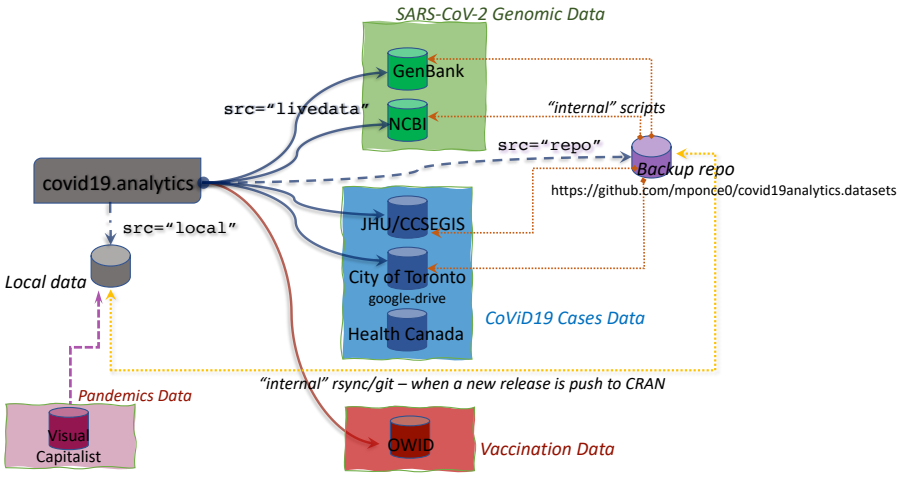

Figure 1: Schematic of the data acquisition flows between the `covid19.analytics` package and the different sources of data. Dark and solid/dashed lines represent API functions provided by the package accessible to the users. Dotted lines are "internal" mechanisms employed by the package to synchronize and update replicas of the data.

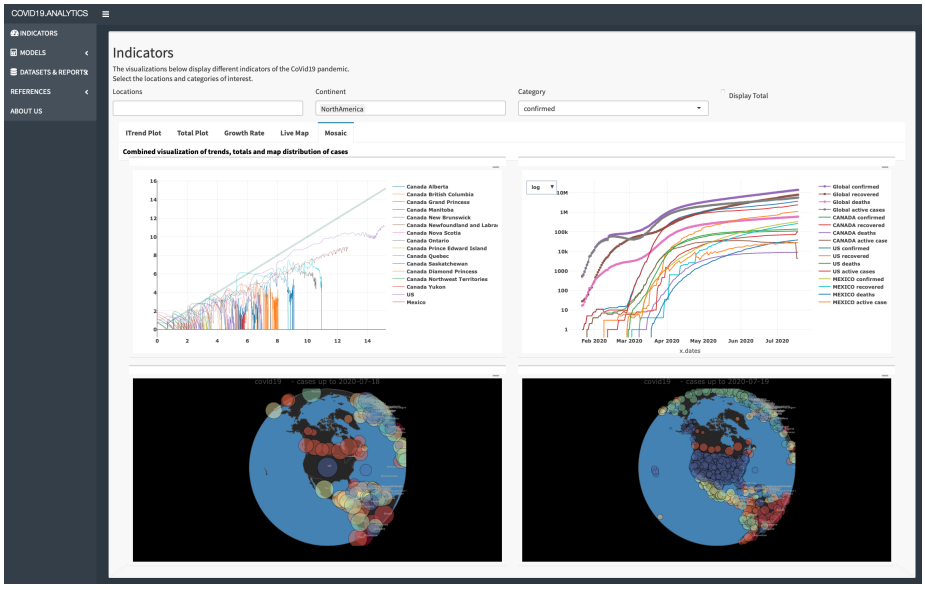

Figure 2: Screenshot of a `covid19.analytics` dashboard implementation – the dashboard can be used through a web-interface or deployed locally in the users' machine.

Moreover, the `covid19.analytics` package offers a modular and versatile approach to the data, by allowing users to input their own data for which most of the package functions can be applied when the data is structured using a *time series* format as described in the package documentation.

The `covid19.analytics` package is also capable of retrieving genomics data, and it does that by incorporating a novel, more reliable and robust way of accessing and designing different pathways to the data sources.

Another unique feature of this package is the ability of incorporating models to estimate the disease spread by using the actual data. Although a simple model, it has shown some interesting results in agreement for certain cases. Of course there are more sophisticated approaches to shred light in analyzing this pandemic; in particular novel *community* approaches have been catalyzed by this too (Luengo-Oroz et al. 2020). However all of these approaches face new challenges as well (Hu et al. 2020), and on that regards counting with a variety, in particular of open source tools and direct access to the data might help on this front.

Further details, documentation, examples and tutorials of the `covid19.analytics` package can be found on the package repository and (Ponce and Sandhel 2020).

Acknowledgments

MP wants to thank all his colleagues at SciNet, especially Daniel Gruner for his continuous and unconditional support, and Marco Saldarriaga who helped us setting up the VM for installing the shiny server hosting the covid19.analytics dashboard web interface.

References

"COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University." 2020. https://github.com/CSSEGISandData/COVID-19.

"COVID-19: Status of Cases in Toronto." 2020. https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/.

"Health Canada – covid19 cases in Canada." 2020. https://health-infobase.canada.ca/src/data/covidLive.

Hu, Yipeng, Joseph Jacob, Geoffrey J. M. Parker, David J. Hawkes, John R. Hurst, and Danail Stoyanov. 2020. "The Challenges of Deploying Artificial Intelligence Models in a Rapidly Evolving Pandemic." *Nature Machine Intelligence* 2 (6): 298–300. https://doi.org/10.1038/s42256-020-0185-2.

Ihaka, Ross, and Robert Gentleman. 1996. "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5 (3): 299–314. https://doi.org/10.1080/10618600.1996.10474713.

Ledford, Heidi, and Richard Van Noorden. 2020. "High-Profile Coronavirus Retractions Raise Concerns About Data Oversight." *Nature* 582 (7811): 160–60. https://doi.org/10.1038/d41586-020-01695-w.

Luengo-Oroz, Miguel, Katherine Hoffmann Pham, Joseph Bullock, Robert Kirkpatrick, Alexandra Luccioni, Sasha Rubel, Cedric Wachholz, et al. 2020. "Artificial Intelligence Cooperation to Support the Global Response to Covid-19." *Nature Machine Intelligence* 2 (6): 295–97. https://doi.org/10.1038/s42256-020-0184-3.

NCBI Resource Coordinators. 2017. "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 46 (D1): D8–D13. https://doi.org/10.1093/nar/gkx1095.

"Open Data Toronto – covid19 cases in the City of Toronto." 2020. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

Our World In Data. 2020. "Our World in Data – CoVid19 Data Repository." https://github.com/owid/covid-19-data/.

Ponce, Marcelo. 2020. *Covid19.analytics: Load and Analyze Live Data from the Covid-19 Pandemic*. https://CRAN.R-project.org/package=covid19.analytics.

Ponce, Marcelo, and Amit Sandhel. 2020. "Covid19.analytics: An R Package to Obtain, Analyze and Visualize Data from the Corona Virus Disease Pandemic." http://arxiv.org/abs/2009.01091.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schriml, Lynn M., Maria Chuvochina, Neil Davies, Emiley A. Eloe-Fadrosh, Robert D. Finn, Philip Hugenholtz, Christopher I. Hunter, et al. 2020. "COVID-19 Pandemic Reveals the Peril of Ignoring Metadata Standards." *Scientific Data* 7 (1): 188. https://doi.org/10.1038/s41597-020-0524-5.

The Comprehensive R Archive Network. n.d. https://cran.r-project.org.

Visual Capitalist. 2020. "Visualizing the History of Pandemics & The Race to Save Lives: Comparing Vaccine Development Timelines." https://www.visualcapitalist.com/history-of-pandemics-deadliest/.