covid19census: U.S. and Italy COVID-19 epidemiolagical data, joined with demographic and health related metrics

Claudio Zanettini
Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
claudio.zanettini@gmail.com

Luigi Marchionni
Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
marchion@jhu.edu
*

Others to be add
Another University
otherstobeadd@example.com

April 26, 2020

Abstract

This is the abstract. Claudio's favourite colour is blue, Claudio's least favirite part of writing a paper is the abstract. Claudio has postponed till the end the writing of the abstract.

Keywords: covid19; R

*Corresponding author; Email: marchion@jhu.edu

1 Introduction

In the mist of a virus pandemic, unraveling the constant flow of epidemiological data is of paramount importance, not only to guide the evaluation and implementation of non-pharmacological interventions (NPI), but also to optimize drug development.

For example, analysis and modeling of COVID-19 confirmed cases and deaths has been employed, in early phases of the pandemic, to assess the effects of NPI in China and Europe [Flaxman et al., 2020, Prem et al., 2020]. More recently, the increased flow of COVID-19 data, and the integration of different sources of information (seasonality of other coronoviruses, U.S. clinical care) has allowed even more long-term predictions of the feasibility and effectiveness of possible containment strategies [Kissler et al., 2020].

Similarly, early evidences of the correlation between Bacille Calmette-Guérin vaccination and COVID-19 outcomes, spur several clinical investigations [Miller et al., 2020, Shet et al., 2020]; WHO]. However, the implications and conclusions of that initial observation were curtailed by subsequent models that included more factors [e.s. age; Fukui et al., 2020]. Overall, these few examples underscore the importance in general, of providing public access to ongoing COVID-19 metrics and in particular, of including multiple heterogeneous collections of data in modeling and analysis, to improve predictions and ultimately address the many challenges of the current pandemic emergency.

The current `R` package provides tools to rapidly extract United States and Italy COVID-19 epidemiological metrics (at county and regional level, respectively) from different sources and combine them with other demographic and health related datasets. The goal of the package is to facilitate multifactorial analysis and modeling of COVID-19 data by the scientific community.

2 Alghorithm

A family of `get` functions is employed by the `R` package to dynamically extract updated time-series data from different on-line sources, combine them and finally return a `dataframe`.

For **U.S** the prefix of the functions to extract data is `getus_`, and it is followed by the specific metric of interest:

- `getus_covid`: extracts data of COVID-19 from the New York Time git repository.
- `getus_dex`: extracts data of DEX, an activity indexes calculated by Victor Couture, Jonathan Dingel, Allison Green, Jessie Handbury, and Kevin Williams based on smartphone movement data provided by `PlaceIQ`.
- `getus_tests`: extract info regarding number of tests performed, their results and hospitalization from the repository of the Covid Tracking Project.

- `getus_all`: executes all the above functions and join the results with other datasets statically contained in the package, and returns a `dataframe` with 304 variables.

Data regarding the household composition, population sex and age and poverty levels (2018), were retrieved from the American Community Survey. Medical conditions, tobacco use, cancer and, data relative to the number of medical and emergency visits (2017) of medicare beneficiaries were obtained from the Mapping Medicare Disparities. The number of hospital beds per county (2020) was calculated from data of the Homeland Infrastructure Foundation.

For **Italy**, the prefix of the function is `getit_` followed by `covid` or `all`.

- `getit_covid`: extracts data of COVID-19 cases, deaths, hospitalizations and tests from the Protezione Civile.
- `getit_all`: executes the above function, join the results with other datasets statically contained in the package and returns a `dataframe` with 64 variables.

Age and sex of the population (2019), first aid and medical guard visits (2018), smoking status (2018), prevalence of chronic conditions (2018), annual-household income (2017), household crowding index (2018) and body-mass index were collect from ISTAT. Prevalence of types of cancer patients (2016), influenza-vaccination coverage (2019) and the number of hospital beds per 1000 people (2017) were obtained from Ministero della Salute. Data of particulate 2.5 (2017) comes from the Istituto Superiore Per La protezione Ambientale.

The package documentation reports and describes each variable (`colnames`) and lists all the data sources of each of the functions. Because of the large amount of variables and, in order to facilitate exploration of the documentation, it was deemed more practical to create separate functions with separate documentation for each of the country.

3 Implementation and use

The package is current available on github. The following code launch the functions and assign the returned `dataframes` to different names.

```
library(covid19census)
dat_it <- getit_all()
```

```
## Italy COVID-19 data up to 2020-04-26 17:00:00 successfully retrived!
```

```
dat_us <- getus_all()
```

```
## US COVID-19 data up to 2020-04-25 successfully retrived!

## US mobility data up to 2020-04-09 successfully retrived!

## US test data up to 2020-04-26 successfully retrived!
```

```
unlist(lapply(list(dat_it, dat_us), class))
```

[1] "data.frame" "data.frame"

Information on the dataframes generated by the two functions are reported in table below [table 1].

	getus_all	getit_all
columns	304	64
counties-regions	2790	21
sources	7	4
from	2020-01-21	2020-02-24

Table 1: Dataframes retuned by the functions. The table reports number of columns, number of unique regions (Italy) and counties (U.S.), unique sources of data was scarped and earliest data related to COVID-19 metrics, of the dataframes returned by the two functions.

Data exploration and modeling can be conveniently performed on a (single) dataframe that contains COVID-19 as well as many other metrics retrieved from multiple sources.

For example, in [figure 1] is reported a correlation analysis of pair of selected U.S. variables.

4 Conclusions

The `R` package `covid19census` extracts and integrates epidemiological COVID-19 data (Italy and U.S at the regional and county level, respectively) with several other demographic and health related indexes. By combining data form different sources, the package is aimed at promoting and simplifying the analysis and modeling of COVID-19 data.

Acknowledgements

Funding : L.M. was supported by NIH-NCI grants P30CA006973, U01CA196390, R01CA200859and the Department of Defense (DoD) office of the Congressionally Directed Medical Research Programs (CDMRP) award W81XWH-16-1-0739.

Conflict of interest : none.

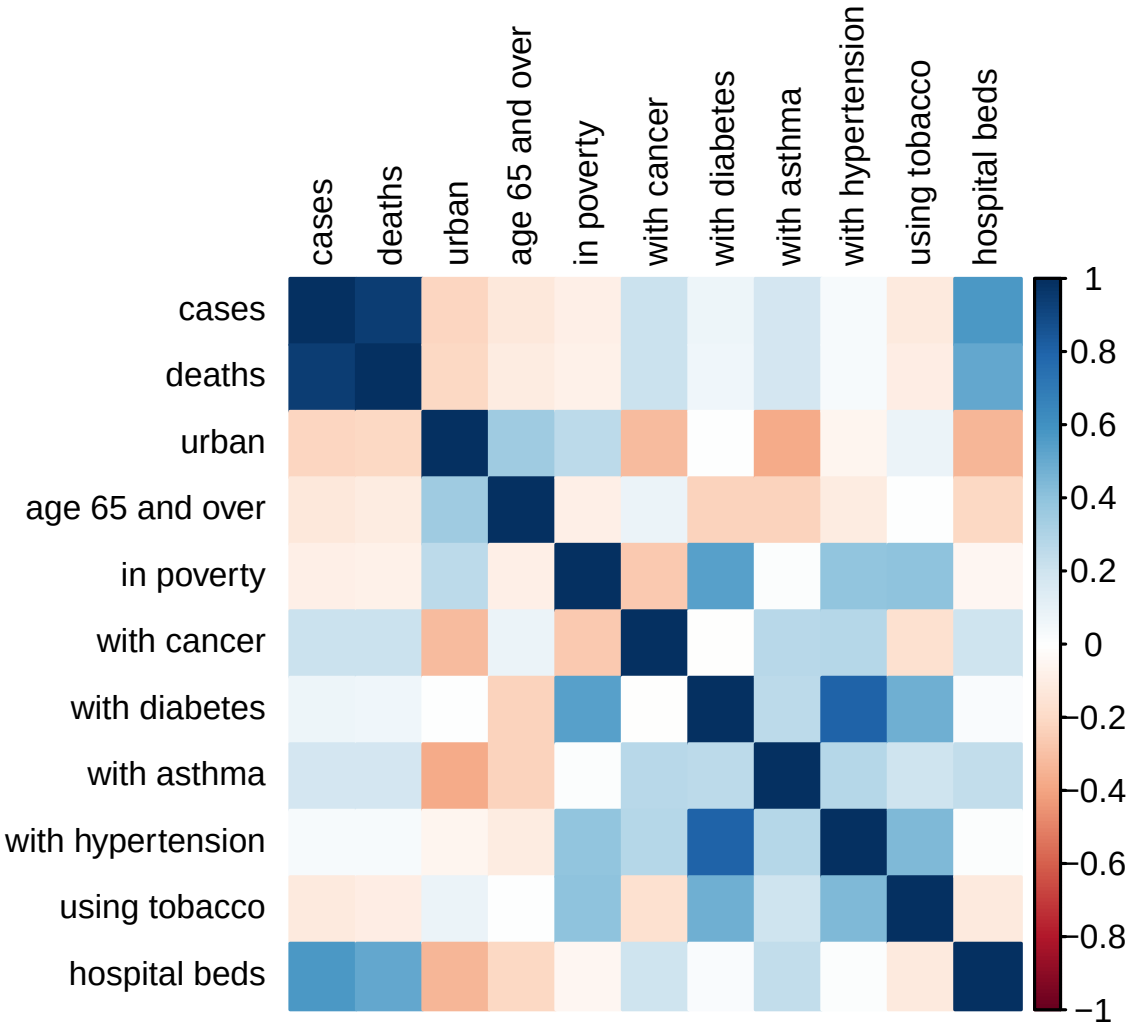

Figure 1: United States Correlation Matrix. An example of exploratory analysis of data from different sources combined by the function 'getus_all'. Colours indicates Pearson's correlation between pairs of variables

References

S. Flaxman, S. Mishra, A. Gandy, H. Unwin, H. Coupland, T. Mellan, H. Zhu, T. Berah, J. Eaton, P. Perez Guzman, N. Schmit, L. Cilloni, K. Ainslie, M. Baguelin, I. Blake, A. Boonyasiri, O. Boyd, L. Cattarino, C. Ciavarella, L. Cooper, Z. Cucunuba Perez, G. Cuomo-Dannenburg, A. Dighe, A. Djaafara, I. Dorigatti, S. Van Elsland, R. Fitzjohn, H. Fu, K. Gaythorpe, L. Geidelberg, N. Grassly, W. Green, T. Hallett, A. Hamlet, W. Hinsley, B. Jeffrey, D. Jorgensen, E. Knock, D. Laydon, G. Nedjati Gilani, P. Nouvellet, K. Parag, I. Siveroni, H. Thompson, R. Verity, E. Volz, C. Walters, H. Wang, Y. Wang, O. Watson, P. Winskill, X. Xi, C. Whittaker, P. Walker, A. Ghani, C. Donnelly, S. Riley, L. Okell, M. Vollmer, N. Ferguson, and S. Bhatt. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Report, March 2020. URL `http://spiral.imperial.ac.uk/handle/10044/1/77731`. Accepted: 2020-03-30T15:10:34Z Publication Title: 35.

Masao Fukui, Kohei Kawaguchi, and Hiroaki Matsuura. Does TB Vaccination Reduce COVID-19 Infection?: No Evidence from a Regression Discontinuity Analysis. *medRxiv*, page 2020.04.13.20064287, April 2020. doi: 10.1101/2020.04.13.20064287. URL `https://www.medrxiv.org/content/10.1101/2020.04.13.20064287v1`. Publisher: Cold Spring Harbor Laboratory Press.

Stephen M. Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H. Grad, and Marc Lipsitch. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science (New York, N.Y.)*, April 2020. ISSN 1095-9203. doi: 10.1126/science.abb5793.

Aaron Miller, Mac Josh Reandelar, Kimberly Fasciglione, Violeta Roumenova, Yan Li, and Gonzalo H. Otazu. Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study. *medRxiv*, page 2020.03.24.20042937, March 2020. doi: 10.1101/2020.03.24.20042937. URL `https://www.medrxiv.org/content/10.1101/2020.03.24.20042937v1`. Publisher: Cold Spring Harbor Laboratory Press.

Kiesha Prem, Yang Liu, Timothy W. Russell, Adam J. Kucharski, Rosalind M. Eggo, Nicholas Davies, Stefan Flasche, Samuel Clifford, Carl A. B. Pearson, James D. Munday, Sam Abbott, Hamish Gibbs, Alicia Rosello, Billy J. Quilty, Thibaut Jombart, Fiona Sun, Charlie Diamond, Amy Gimma, Kevin van Zandvoort, Sebastian Funk, Christopher I. Jarvis, W. John Edmunds, Nikos I. Bosse, Joel Hellewell, Mark Jit, and Petra Klepac. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet*

Public Health, 0(0), March 2020. ISSN 2468-2667. doi: 10.1016/S2468-2667(20)30073-6. URL https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(20)30073-6/abstract. Publisher: Elsevier.

Anita Shet, Debashree Ray, Neelika Malavige, Mathuram Santosham, and Naor Bar-Zeev. Differential COVID-19-attributable mortality and BCG vaccine use in countries. *medRxiv*, page 2020.04.01.20049478, April 2020. doi: 10.1101/2020.04.01.20049478. URL https://www.medrxiv.org/content/10.1101/2020.04.01.20049478v1. Publisher: Cold Spring Harbor Laboratory Press.