covid19census: U.S. and Italy COVID-19 epidemiolagical data 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
otherst@example.com

April 25, 2020

Abstract

This is the abstract.

It consists of two paragraphs.

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 implementation and evaluation of non-pharmacological interventions, but also to optimize drug development.

- Examples of epdidemiological alone or + dother data guiding interventions: [Kissler et al., 2020]: proposed non pharmacological intervention

[Wu et al., 2020]: correlation p2.5
BCG: clinical trial on BCG
We need data banks, repositories of aggregated data

- Examples of that and databanks of epidemiological as well as genetic data. The traking project

- Examples of R package ccdcovidview

Boom our package

2 Alghorithm

A family of `get` functions is employed by the `R` package to extract updated time-series data dynamically from different on-line sources and combine them and 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 and 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, instead of creating a single function with an argument relative to 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 2020-04-25 17:00:00 successfully retrived!
```

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

```
## US COVID-19 data up to 2020-04-24 successfully retrived!

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

## US test data up to 2020-04-24 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 (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.

Therefore, data exploration and modeling of COVID-19 metrics can be conveniently performed on single dataframes that combine heterogeneous datasets from multiple sources.

For example, in (1) correlation analysis of pair of selected U.S. variables and relative visualization.

4 Discussion

We are cool, the other people suck

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.

References

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

Xiao Wu, Rachel C. Nethery, Benjamin M. Sabath, Danielle Braun, and Francesca Dominici. Exposure to air pollution and COVID-19 mortality in the United States.

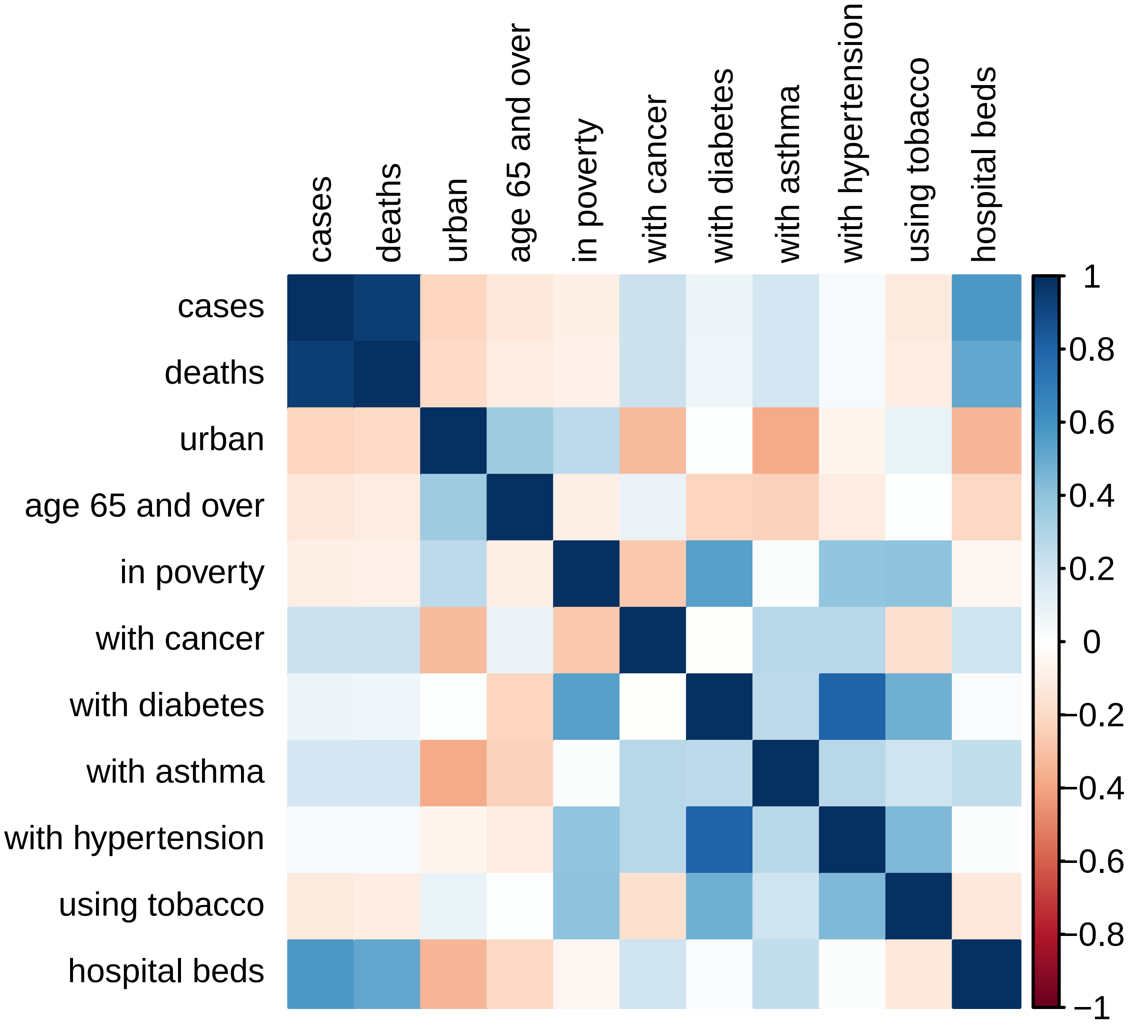

Figure 1: United States Correlation Matrix. An example of exploratory analysis of data from different sources combined by the function ‘getus_all‘. Colours indicates Pearson’s correlation between pairs of variables

medRxiv, page 2020.04.05.20054502, April 2020. doi: 10.1101/2020.04.05.20054502. URL https://www.medrxiv.org/content/10.1101/2020.04.05.20054502v1. Publisher: Cold Spring Harbor Laboratory Press.