An R package and a website with real-time data on the COVID-19 coronavirus outbreak

Tianzhi Wu[1], Xijin Ge[2,*], Guangchuang Yu[1,*]

[1]Department of Bioinformatics, School of Basic Medical Science, Southern Medical University, Guangzhou, China
[2]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA.

Corresponding: Guangchuang Yu (gcyu1@smu.edu.cn); Xijin Ge (Xijin.Ge@sdsate.edu)

Keywords: coronavirus, SARS-COV-2, 2019-nCoV, COVID-19

To provide convenient access to epidemiological data on the coronavirus outbreak, we developed an R package, nCov2019 (https://github.com/GuangchuangYu/nCov2019). Besides detailed real-time statistics, it also includes historical data in China, down to the city-level. We also developed a website (http://www.bcloud.org/e/) with interactive plots and simple time-series forecasts. These analytics tools could be useful in informing the public and studying how this and similar viruses spread in populous countries.

Introduction

The current outbreak of the novel coronavirus, SARS-COV-2 [1], originated at the end of 2019 from Wuhan, a city in central China with a population of about 11 million. According to a recent update on February 19, 2020, from the China National Health Commission (CNHC), confirmed cases across 17 countries reached 75,726, 98% of which is in China [2]. World Health Organization (WHO) declared it a global health emergency. The SARS-COV-2 outbreak is still spreading internationally and has had an enormous impact on human health, social life, and the economy. It also puts considerable stress on the healthcare system in China.

As the outbreak is still escalating rapidly, researchers need to gather the epidemiological data to study the spread of the disease. Our goal is to create an R package that can facilitate the dissemination of accurate data about this outbreak. The number of confirmed cases is updated daily by the CNHC and similar agencies in other countries. We have created the nCov2019 package that integrates the up-to-date public information of the infection based on API (application programming interface) to the Tencent SARS-COV-2 news website[3]. The package also offers historical data collected by a non-governmental organization Dingxiangyuan [4].

This kind of detailed historical data for dozens of provinces and hundreds of cities in China provides a unique opportunity to study how the virus spreads in a vast, populous country. The hundreds of Chinese cities could also be considered as independent outbreaks, as many

of them are far from the epicenter, and many are effectively on lockdown from the end of January 2020.

Query the latest statistics data

Users could utilize the remotes package to install it directly from GitHub. See supplementary document 1 for a more detailed tutorial. Once the nCov2019 package was loaded, we can get the latest statistics data in just one command: *x <- get_nCov2019(lang='en')*. For ease of use, the package supports two languages, Chinese (default) and English.

The primary function is to get real-time data for each geographical area. Since more than 98% of confirmed cases concentrated in China, researchers may more concerned about the details of the infection in China. Printing the object *x* will directly show summary information of the total amount of confirmed cases in China and the last updated time. We re-defined the [operator to extract data. When no region is specified, *x[]* will return a more intuitive statistical data for China, including cumulative diagnoses, mortality, cure rate, *etc.* for each province. Users could obtain a more granular scale data, for example, *x['Hubei']* or *x['湖北']*, which depend on your language setting, will return a data frame contains the latest data available for each city of the selected province. And using *x ['global',]* will return another data frame for the global landscape view of each country. Also, using the parameter *by='today'*, will return the newly added number of the current day, such as freshly infected cases.

Get the historical statistics data

We organize and summarize daily data at province and city levels. The information was updated daily and can be loaded into R using the *x <- load_nCov2019()* function. Epidemiologists will find the data useful for modeling and forecasting. The methods we provided to access the data is consistent by using the [operator and summary function. Using the command *x[]* will return the historical information of all cities in China, and if any province is specified, e.g., *x['Hubei']*, the historical data of all cities belong to the selected region will be returned. Users can use *summary(x)* to get summary data at the province level. If a province is specified, e.g., *summary(x, "Guangdong")*, it will only return summary data of the selected region. See supplementary document 1 for a more detailed tutorial.

Plot the Geographic maps

Geographic visualization is an effective way to observe the spatial patterns of disease spread. Presenting the spread of the disease on the map is supported in the nCoV2019 package. Users can use the command, *plot(x)*, to visualize disease information on the world map. If users want to plot data on a China map at the province or city level, they need to provide GIS files. Plotting data on a selected province or geographical region is also supported if the GIS information was available.

In combination with R package chinamap [5], it is easy to visualize data on the Chinese map at the province level (see Figure 1). Using historical data, users can visualize data of specific dates on a map. For example, *plot(x, region='china', chinamap=cn, date='2020-02-01')* will plot

historical data on 2020-02-01. With historical data and the ability to visualize data on a map, we can incorporate temporal and spatial information to create an animation to show how SARS-Cov-2 is spreading.

Interactive visualization

To enable users to access data about this outbreak easily, we created an interactive website based on the RStudio Shiny framework. This web app is available in both English (http://www.bcloud.org/e/) and Chinese (http://www.bcloud.org/v/). See a detailed demonstration in supplementary document 2. It displays various statistics in geological layers, ranging from the world, country (China), provinces, cities. For example, Figure 2 shows detailed infection statistics of all provinces except Hubei, the epicenter with most of the cases. The sharp, exponential growth of coronavirus cases observed in Wuhan, Hubei, is not seen in other provinces. Instead, the total confirmed cases outside Wuhan are stabilizing as of February 23, 2020. The extreme measures of locking down tens of millions of people that the Chinese government took are working.

Users can interactively select their regions of interest and check both the historical and real-time data. This app also has a simple forecast for total confirmed cases and deaths. We first converted the number of confirmed cases in log-scale as a time series data. We used the *ets* function in the R package *forecast* with default settings, which automatically selects error type, trend type, and season type. We also converted the raw number of cases as percent changes compared to the previous day and conducted a similar forecast.

Discussion

Real-time data collection is essential not only to inform the public but also for scientists and health officials. Thanks to the cooperation and open source, we can regularly organize public data and develop this package. Our nCov2019 package aims to reduce the difficulties for clinicians and other researchers in obtaining comprehensive data. Also, we developed a web application for researchers to conduct interactive analytics.

Currently detailed, city-level data is only available for mainland China. These data sources occasionally will change data format, which requires us to closely monitoring this package. If the APIs stopped providing data, then the real-time data will not be updated. But the historical data will remain accessible.

References:

[1]. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, China Novel Coronavirus I, Research T (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 382 (8):727-733. doi:10.1056/NEJMoa2001017

[2]. http://www.nhc.gov.cn/xcs/yqfkdt/202002/4a1b1ec6c03548099de1c3aa935d04fd.shtml

[3]. https://news.qq.com/zt2020/page/feiyan.htm

[4]. https://ncov.dxy.cn/ncovh5/view/pneumonia

[5]. https://github.com/GuangchuangYu/chinamap

[6]. Hyndman, R.J., Koehler, A.B., Ord, J.K., and Snyder, R.D. (2008) Forecasting with exponential smoothing: the state space approach, Springer-Verlag.

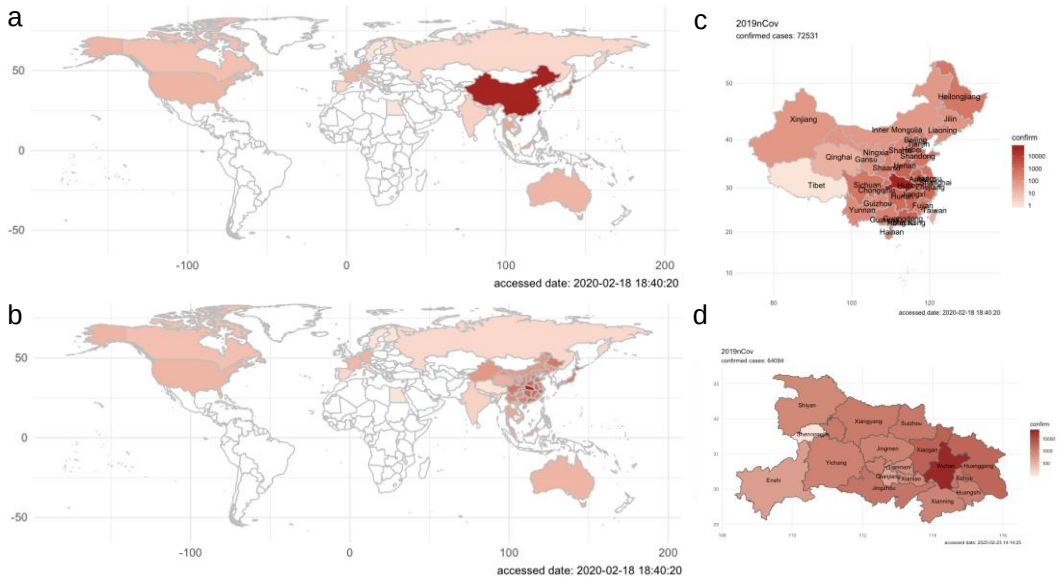

Figure 1. a). and b). Geographic visualization for tracking the global spreading trend. c). and d). Geographic map with further details in China region, Chinese map, and city-level map in Hubei province, China, respectively.

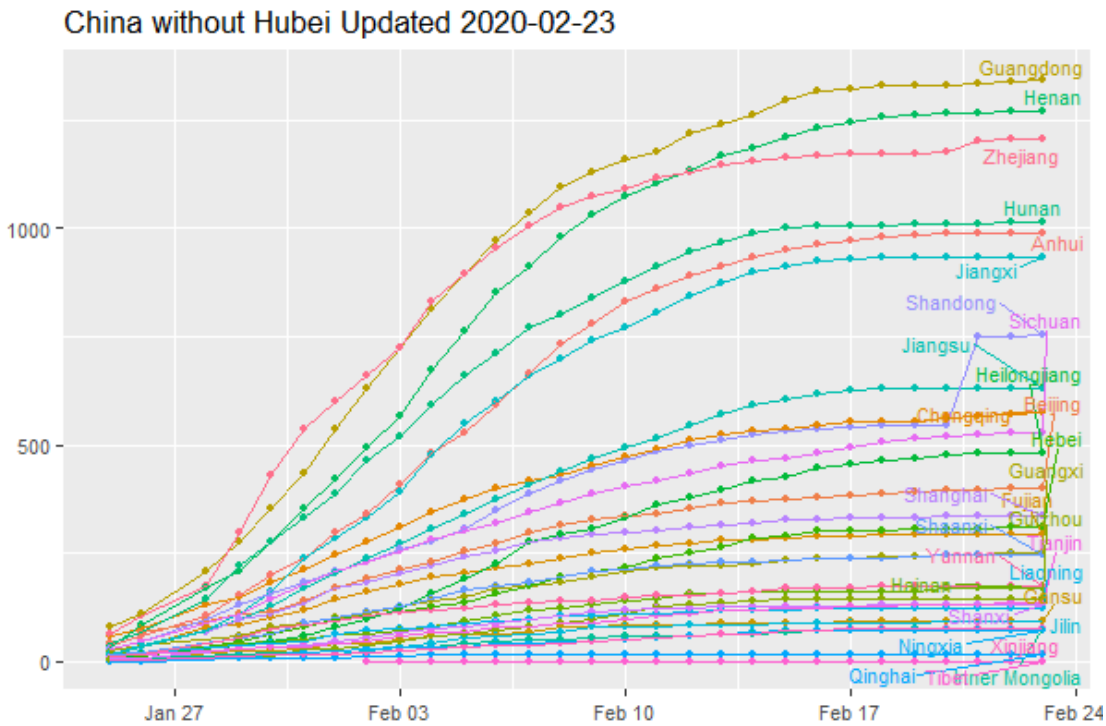

Figure 2. Through the interactive website, users can get detailed statistics for all provinces in China. The epicenter, Hubei province, was excluded. Similar plots can be obtained easily for hundreds of cities across China using the website.