Data Sources for "Forecasting the Spread of COVID-19 under Different Reopening Strategies"*

Meng Liu Raphael Thomadsen Song Yao

Olin Business School
Washington University in St. Louis

This draft: June 3, 2020

*Please contact Liu (mengl@wustl.edu), Thomadsen (thomadsen@wustl.edu), or Yao (songyao@wustl.edu) for correspondence.

In this online document, we detail our data sources for the manuscript "Forecasting the Spread of COVID-19 under Different Reopening."

Data

Our data come from a multitude of sources. We lay out the sources for each of these in turn. The final dataset for estimation can be downloaded at `https://github.com/songyao21/covid_data_depot`. The only exception is that we are not allowed to share data for our social distancing metric, which comes from SafeGraph. However, SafeGraph is allowing academic COVID-19 researchers to get access to the data for free by signing up at `https://www.safegraph.com/covid-19-data-consortium` (accessed June 2, 2020).

Positive Cases

Data of positive cases are based on the COVID-19 data published by the New York Times (`https://github.com/nytimes/covid-19-data`, accessed on May 17, 2020). The data contain the daily confirmed case counts for 2,953 U.S. counties or county-equivalents. We exclude cases in the states of New York, New Jersey, and Connecticut due to the large outbreak there and the complicated relationship between New York City (which is the seat of 5 counties) and the surrounding counties. We also drop 3 of the remaining counties because we do not have social distancing data for 2 of them, and we cannot match the demographic data for a third (Oglala Lakota County, SD). This reduces the number of counties to 2,864. Finally, we remove counties that had no confimed cases during our sample period of Feb 1, 2020 to May 18, 2020. After all the above-metioned filters, we have a panel of 2,834 counties. These counties account for 89.8% of the total U.S. population and 66.97% of the total U.S. confirmed cases till May 29, 2020.

There are a few days where there are negative cases that are reported. These are generally corrections to previous over-reporting. Thus, we clean the negative numbers of cases by subtracting the absolute value of the negative cases from the proceeding day. In the event that this procedure leads to a negative number of the proceeding day, we iterate again.

Social Distancing

We use social distancing data from the company SafeGraph, which collects cellphone GPS data from U.S. residents, and has made them available for free to academics studying COVID-19. These data are collected through a series of pings that the company receives for all users who have installed a number of smartphone apps. The list of apps that collect this information is kept as a trade secret. For each county, we use the fraction of cellphones that stayed near home for the whole day as our measure of social distancing. The SafeGraph data are published at the Census Block Group level. To accommodate other data sources which are available at a less granular level, we aggregate the this variable to the county level by taking the weighted median, using the number of cellphones in each Census Block Group as the weight.

	(1) N	(2) Mean	(3) Std Dev.	(4) Min	(5) Max
Reported cumulative cases	125,389	199.40	1,142.00	0.00	63,690.00
Reported new cases	125,389	3.48	8.03	0.00	145.10
Average temperature (in Celsius)	125,389	12.90	6.72	-16.40	32.06
Rain (mm)	125,389	3.48	8.03	0.00	145.09
Humidity (percentage point)	125,389	67.08	15.31	8.75	100.00
Social_distancing (home ratio)	125,389	0.33	0.07	0.07	0.83

Table A1: **Summary Statistics**

Weather data

We gathered historical daily rain and temperature data for the period from February 1 to May 18, 2020 from the Global Historical Climatology Network of National Oceanic and Atmospheric Administration (NOAA) (source: `https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html`, accessed on May 21, 2020). The data on humidity is obtained from the U.S. Local Climatoligical Data of NOAA (source: `\url{https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html},accessedonMay25,2020`

These raw weather data are at their respective weather station level. For each county, we match the weather stations that are within 50 miles from the population centroid of the county. We use the average weather data across the matched stations as the weather of the county. For a small number of counties where no matched stations are found, we use the daily averages of the state the county is in to impute.

Putting it all together

Our sample is an unbalanced panel because counties start to have positive number of confirmed cases on different dates. The earliest date we observe in the sample is Jan 29, 2020, and the last day is May 29, 2020. However, because our weather data cover a shorter period of time, our final sample for estimation is February 1 to May 18, 2020. Note that we construct actual cases using reported cases 5 days later, and thus the corresponding sample period of reported cases is Feb 6, 2020 to May 23, 2020.

Summary statistics of all of the variables we use in the estimation are presented in Table A1. Note that our case data proceed past the dates used for estimating the model and are up to May 29, 2020. We use those data for validating the model. Those data are publicly available, but we are happy to supply summary statistics for this hold-out sample upon request.