

# Data Sources for “Forecasting the Spread of COVID-19 under Different Reopening Strategies”<sup>\*</sup>

Meng Liu   Raphael Thomadsen   Song Yao

Olin Business School  
Washington University in St. Louis

This draft: June 26, 2020

---

<sup>\*</sup>Please contact Liu ([mengl@wustl.edu](mailto:mengl@wustl.edu)), Thomadsen ([thomadsen@wustl.edu](mailto:thomadsen@wustl.edu)), or Yao ([songyao@wustl.edu](mailto: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](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 June 24, 2020). The data contain the daily confirmed case counts for 2,953 U.S. counties or county-equivalents. The case data for the five boroughs of New York City, however, are not recorded separately by New York Times. In this case, we use the data published by the Health Department of New York City in lieu of the five boroughs (<https://github.com/nychealth/coronavirus-data>, accessed on June 24, 2020). The case data of Kansas City, Missouri are also recorded separately because it overlaps with 4 adjacent counties. We attribute the cases of Kansas City to Jackson County, Missouri because most of the city lies within Jackson County. We also drop 3 counties because we do not have social distancing data for 2 of them, and we cannot match the population data for a third (Oglala Lakota County, SD). Finally, we remove counties that had no confirmed cases during our estimation sample period of Feb 1, 2020 to May 23, 2020. After all the above-mentioned filters, we have a panel of 2,925 counties. These counties account for 99.76% of the US population and 99.91% of the total U.S. confirmed COVID-19 cases till June 23, 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.

	(1) N	(2) Mean	(3) Std Dev.	(4) Min	(5) Max
Reported new cases	131,344	113.95	643.84	0	21,740
Average temperature (in Celsius)	131,344	12.65	6.73	-16.40	32.06
Rain (mm)	131,344	3.47	7.94	0.00	145.09
Humidity (percentage point)	131,344	66.94	15.25	8.75	100.00
Social_distancing	131,344	0	1.59	-18.92	5.51
Population (in Thousand)	2,925	111.87	344.46	0.63	10039.11

Table A1: **Summary Statistics**

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.

## 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 Climatological Data of NOAA (source: <https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html>, accessed on May 25, 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 June 23, 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 June 23, 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.