**Data description**

Qing Yao [(qy2290@columbia.edu)](mailto:(qy2290@columbia.edu))

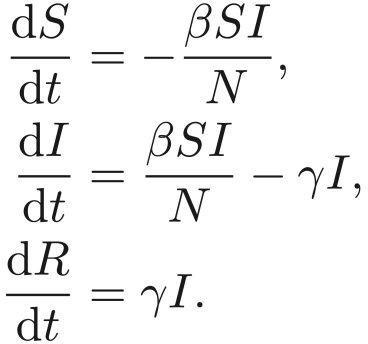
June 11 2023

**- Aim: model the 2020 The third wave, infer the beta of SIR**

**1. Model description**

In the simple SIR model

We have



wpsoffice

* S(t) represents the population susceptible to, but not yet infected with the disease.
* I(t) refers to the number of currently infectious individuals
* R(t) indicates those who have recovered and now possess immunity.
* Βeta describes the effective contact rate of the disease. We assume the beta is the a function of various features X, such as: mobility, social-economic characteristics, weather conditions (humidity), access to public transportations.
* We would like to infer a explicit expression of the function f. Some Xs are changing with t (daily) while others remain stable within our chosen time frame.
* Gamma is the mean recovery rate: that is, 1/γ is the mean period of time during which an infected individual can pass it on, we use 5 as the 1/gamma.

**2. USA Geography basics**

County level:

* In the latest Census statistics, there are **3143** counties in the 50 states (stars) in the US. (details are not included in here).
* https://www.census.gov/programs-surveys/geography/about/glossary.html#
* The **fips ≡** **geoid**: The first two digits represent state code, the last three present the county code inside the state/area.
* There territories, also having county codes: Puerto Rico(72xxx, 78 in total), US Virgin Island (78xxx, 3 in total), Northern Mariana Islands (69xxx, 4 in total), American Samoa (60xxx, 5 in total), Guam (66xxx, one in total). But they are not sovereign entities (https://en.wikipedia.org/wiki/Territories\_of\_the\_United\_States)
* In our study, we only look at the counties in of 48 states in the mainland, the total number is **3108. The Hawaii and Alaska are not included. See the file: fips\_mainland.csv**
* Note, this number is updated in 2019. However, some datasets are collected before 2019, I tried to align them to 3108/3143. (Example, some counties merged into two, some county separated into two, some counties changed name and county code, the details are described in the Appendix.)

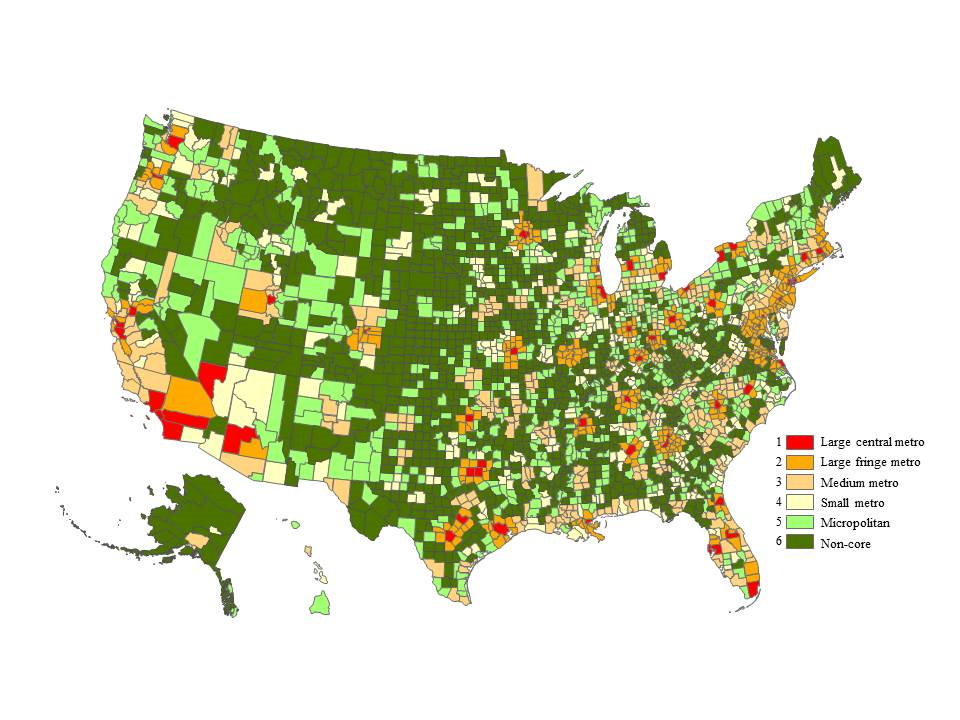


Figure1 The urban rural county codes the data is contained in urban\_rural\_codes.csv

**3. Covid19 confirmed cases**

Figure plots the daily confirmed case, which is roughly the dI/dt \* observation rate

This rate is also called *reporting rate* and describes the fraction of people who tested positive or get infected reports their infection to the government.

Note: The reporting rate varies with times when the wave starts, from 3-4% increases linearly to 20-30%

(Q: what should we use for the third wave? Maybe similar to the first wave as linear increase to 30% and then be steady?

What we )

**ts\_covid19\_confirmed\_US\_mainland.csv**

This dataset is the accumulated number of reported infections (cases), when you take diff(ts), it will give you the figure above.

*https://github.com/CSSEGISandData/COVID-19/tree/master*

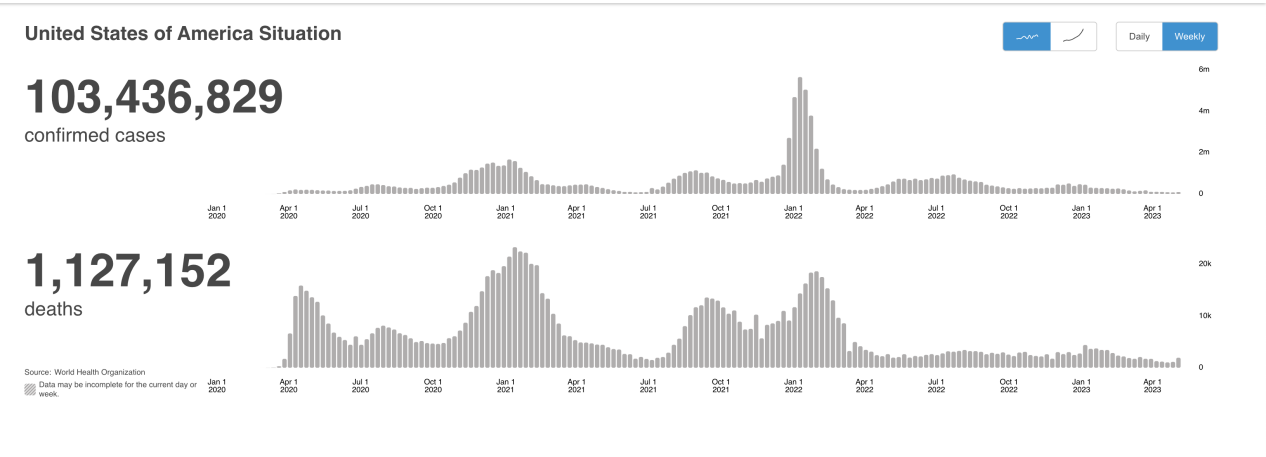


Figure2 The number of daily new cases

There are unassigned data for each state. I deleted them.

**4. Variables for counties**

**4.1 Mobility data**

We possess daily records detailing people's visits to different locations, including their home addresses. We then compile this data into daily statistics, showing movements from one county to another. Please note that this data is noisy and reflects limited inter-county commuting. We've classified Points of Interest (POIs) into seven categories. Our analysis focuses on the number of people visiting each category in (or go to) each county, characterizing this as a time series to depict travel patterns. I compressed into a zip file **ts\_POI.zip.**

**2020\_c-{}.csv** .format(0,1...6) the starting day of this is from Jan, 06, 2020

**2021\_c-{}.csv** .format(0,1...6)

(https://docs.safegraph.com/docs/places)

The classification used in safegraph is ‘We currently reference the **2017** version of NAICS’ there are some changes in 2022

This set of data probably be standardised by the population

Some of the categories like 0 does not have visits for all the fips, they are all zeros.

**4.2 Demographic & socioeconomic**

a) population, population density, age groups (10 years? 0-9, 10-19, …, 70-79, 80+), Black, Hispanic, Asian, median household income, bachelor’s degree, household size, employment, urban/rural.

**demographic\_socialeco\_2021.csv**

**urban\_rural\_codes\_2013.csv**

**land\_density\_2020.csv**

Data originals: 2017-2021 5-year ACS, Changed GEOID to FIPS. The description of variables can be checked at

https://walker-data.com/tidycensus/articles/basic-usage.html

Some variable description:

The urban-rural code, 1 stands for the large metro, 6 means non-core (<https://www.cdc.gov/nchs/data_access/urban_rural.htm>)

The population density is Population per square mile and is found in <https://covid19.census.gov/datasets/average-household-size-and-population-density-county/explore?location=22.847890%2C50.195024%2C2.10>

**4.3 Meteorological (time varying)**:

temperature, specific humidity, total precipitation, climate\_zones

**climate\_zones.csv**

Temperature**:**

**ts\_weighted\_area\_tmean\_2020.csv**

**ts\_weighted\_area\_tmean\_2021.csv**

Total precipitation

**ts\_weighted\_area\_ppt\_2020.csv**

**ts\_weighted\_area\_ppt\_2021.csv**

specific\_humidity

1000hpa (roughly 0m/just above the ground)

**ts\_weighted\_area\_sh\_2020\_ip.csv**

**ts\_weighted\_area\_sh\_2021\_ip.csv**

When you open the datafile using csv it maybe not show all the data, just use python or load the data directly.

Temperature mean and total precipitation is from:

<https://github.com/rmp15/PRISM-grids-into-FIPS-ZIP-censustract-USA/tree/main>

Tmean is from Jan 01 2020 to Dec 31 2021, daily, and averaged through all the tmean of inside one county, is in degree.

Ppt is from Jan 01 2020 to Dec 31 2021, daily, and add up all the ppt of fips inside one county, is mm.

Specific humidity is from:

This parameter is the mass of water vapour per kilogram of moist air. The total mass of moist air is the sum of the dry air, water vapour, cloud liquid, cloud ice, rain and falling snow. Some of the literature will use relative humidity, the relative humidity is (This parameter is the water vapour pressure as a percentage of the value at which the air becomes saturated (the point at which water vapour begins to condense into liquid water or deposition into ice). For temperatures over 0°C (273.15 K) it is calculated for saturation over water. At temperatures below -23°C it is calculated for saturation over ice. Between -23°C and 0°C this parameter is calculated by interpolating between the ice and water values using a quadratic function.)

<https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview>

The resolution is 0.25° x 0.25°, roughly cover the areas of 707 km^2. However, some counties are small (smaller than this area) and then can not be masked from this resolution. Interpolate the data before applying the mask, after interpolating, Manassas Park, Virginia, fips 51685, is tiny (7.86 km^2), using the data of Prince William County, Virginia (fips 51153) as the former one is inside the later one.

**4.4 Political (Voting)**:

**voting\_data.csv**

Source: <https://uselectionatlas.org/BOTTOM/store_data.php>

I deleted the overseas voting and the statistics of districts

The term "county" is used in 48 US states, while Louisiana and Alaska have functionally equivalent subdivisions called parishes and boroughs respectively.

In Virginia, the statistic level is city

**4.5** **Vaccination**

Disease outcomes (time varying): COVID

**ts\_vaccinations.csv**

https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

https://github.com/CSSEGISandData/COVID-19/tree/master/csse\_covid\_19\_data/csse\_covid\_19\_time\_series

The data uses 2261, to align the length of the 3143, I duplicated the 2261 data and renamed them as 2063 and 2066. (I assume we use the percentages of completeness, then the absolute values are not important. If we do need absolute value, we can distributed the quantity of 2261 by the populations of 2063 and 2066.)

The length of the dates of each county is not the same (320 or 319) only 1207 counties has the data of 09/11/2021. This is the starting date of the records. Therefore, we can delete the data of this date.

**4.6 NPI (time varying)**:

Policy indices:

overall government response index (all indicators)

containment and health index (all C and H indicators)

stringency index (all C indicators, plus H1 which records public information campaigns)

economic support index (all E indicators)

**ts\_OxCGRT\_USA\_differentiated\_withnotes\_2020.csv**

**ts\_OxCGRT\_USA\_differentiated\_withnotes\_2021.csv**

Sources: <https://github.com/OxCGRT/covid-policy-tracker>

Note to cite the data

Recommended reference for academic publications: “A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker).” Nature Human Behaviour. https://doi.org/10.1038/s41562-021-01079-8

The data is in the state level, I deleted a state called Washington DC, there existing a state called Washington. I added state number code

**Appendix**

County changes

* Shannon County, SD (FIPS code = 46113) was renamed Oglala Lakota County and assigned anew FIPS code (46102) effective in 2014
* 02261 Valdez-Cordova Census Area ZIP Code separted into 2063 (Chugach Census Area) and 2066 (Copper River Census Area) in 2019
* Fips code 2158 (Kusilvak Census Area) previously as 02270 (Alaska      Wade Hampton Census Area)
* 51515 Virginia   Bedford city is Virginia, 2013: The independent city of Bedford (FIPS 51515) merges into Bedford County (FIPS 51019).
* Hoonah-Angoon Census Area, AK (FIPS code = 02105). In 2007, Skagway-HoonahAngoon Census Area (FIPS code = 02232) was split into Hoonah-Angoon Census Area and Skagway Municipality (FIPS code = 02230).
* In 2008, Prince of WalesHyder Census Area was created from the remainder of the former Prince of Wales-Outer Ketchikan Census Area (FIPS code = 02201) after part (Outer Ketchikan) was annexed by Ketchikan Gateway Borough (FIPS code = 02130 and another part was included in the new Wrangell Borough.
* Wrangell City and Borough, AK (FIPS code = 02275).In 2008, Wrangell City and Borough was created from part of Wrangell-Petersburg Census Area (FIPS code = 02280) and part of Prince of Wales-Outer Ketchikan Census Area (FIPS code = 02201). This entity has a category code in the 2013 and 2006 NCHS schemes, but not on the 1990 census-based scheme.
* Clifton Forge City, Virginia (FIPS code = 51560). In 2001, Clifton Forge city, Virginia, formerly an independent city, merged with Alleghany County (FIPS code=51005). This entity has a category code in all three of the NCHS schemes. The category codes for Clifton Forge City are the same as those for Alleghany County.