**COVID-19 Location Data Use Case**

The data consists of two data tables.

1. Patient Trajectory (Route) Information.

* patient\_id: patient level identifier
  + note that any given patient may have anywhere from 1 to 40+ location tuples in their trajectory
* date: date of observed location (at the day level)
* province (of the location)
* city (of the location)
* type: the type of location
  + airport, hospital, store, public transportation
* latitude
* longitude

1. Patient Demographic Information

* patient\_id: patient level identifier
* sex: male or female
* age: age bins, *e.g.,* 50s, 40s, etc.
* country (of the patient)
* province (of the patient)
* city (of the patient)
* infection\_case: means of infection, *e.g.*, overseas inflow, contact with patient, etc.
* symptom onset date
* confirmed date
* released date
* deceased date
* state: released, isolated, deceased

Missing value counts:

patient\_id 0

date 0

province\_x 0

city\_x 0

type 0

latitude 0

longitude 0

sex 1319

age 1319

country 0

province\_y 0

city\_y 144

infection\_case 337

infected\_by 6203

contact\_number 5677

symptom\_onset\_date 6500

confirmed\_date 0

released\_date 5493

deceased\_date 8075

state 0

Due to high variance in trajectory lengths, we drop the patient identifier. If we tried to model things on a patient level we would have a very large, sparse data set which is not the focus of the proposed synthesis method.

\*\***We have individual level data over time. So if we drop time and model counts, we assume each points corresponds to a different person. The counts would be a function of how many locations were observed for a given individual (the model results wouldn’t actually be meaningful).**

**\*\*We are going to explicitly state that we assume each location corresponds to one individual. This is not a realistic treatment of the original data, but it allows us to make a good illustration of the abilities of the synthesis method to generate synthetic data for general use cases.**

We select date, latitude, longitude, sex, age, and state. For simplicity, we drop observations with missing values for sex and age (these could be imputed).

We want to consider two use cases.

1. Very granular data. This is the form in which the data originally exists. What is a use case for this? Could be useful to show that additional effort for synthesization is useful here, *e.g.*, including date.
2. Aggregated data. Consider the data in various forms of aggregation and show that a model for death rates achieves similar results across aggregation levels. Our synthesized data is useful for more than one thing!
   1. Compare estimation of disparities in rates between urban and rural areas and sex

\*\* need to convert only the last point of deceased patients’ trajectories. Just running the model now to see if we get anything interesting.

\*\* a shortcoming of the actual results could be the data collection process. Are individuals with COVID-19 from rural parts of the country as likely to show up as people in highly populated areas? If not, how would this bias the results?

**The R function we are using returns various levels of aggregation – road, quarter, city, province, ISO3166-2-lvl4, postcode, country, country\_code.**

**We can apply models at multiple levels of aggregation. Use models to perform different hypothesis tests, such as disparities in urban/rural death rates or death rates by sex.**

**\*\*\*What if we did the modeling from the “perspective” of the locations? If we added a temporal aspect then we would be counting the number of COVID-19 positive individuals at each location on each day.\*\*\* I think this would be a more valid use of the data??? And would lend itself to using the granular data as a baseline and then aggregating to different levels to compare results.\*\*\***

**\*\*\*Let’s continue treating each point as a unique person. This is how the rest of the paper is written, and we will continue in that theme.\*\*\***

**Use case 1: granular, temporal locations. Sending alerts to nearby individuals, or disease mapping (covid-19 map shown in paper). What is the proximity of real points to synthetic?**

**Use case 2: multiple levels of aggregation. Consider post code and ISO level. Can we get good death rate estimates.**

**We get the above from one data set!**