**Thesis project: Assessing correlation between changes in self-reported and cell-phone geolocated mobility using two large datasets**

**Background:**​

* Increasing use of quantifying human movement during the pandemic including the more novel cell-phone geolocation​
* Important to ground-truth passively collected from cell phone pings​

**Data:**​

* DELPHI for self-reported movement (~50K samples per day in US from facebook)​
* Safegraph for mobility (CBG to POI weekly visitor flows)​

**Method:**​

* Use spatial regression to assess correlation between county-level changes in self-reported movement to places such as school, grocery store and restaurant and mobility flows to the same categories of places​
* Geographical region: Georgia

**Datasets:**

[Safegraph:](https://docs.safegraph.com/docs)

* We have access to the “Weekly Patterns” Data
  + Specifically, the Georgia data has been downloaded and somewhat processed in a more usable format
  + Dataset line-listed by “Points of Interest” (POIs)
  + POIs can be categorized/labelled by type (ex. Shop, school, café/restaurant etc)
    - Erica would need to come up with a method to do this
    - POIs are labeled by [NAICS](https://www.census.gov/naics/) code
  + Data shared with Erica in Sharepoint under “GA\_Safegraph\_Weeklydata”
    - Can consider expanding geography or timeline, but would need extra steps to process data
* Relevant literature:
  + [Safegraph’s internal documentation of variables](https://docs.safegraph.com/docs)
  + SafeGraph “colab” notes on specific topics:
    - [Quantifying sampling bias](https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ#sandboxMode=true&scrollTo=LxBDP8E_uqO6)
    - [Longitudinal analysis](https://colab.research.google.com/drive/16BELpcum4TKoH-5wg8Xym_CGgIGgpu1I?usp=sharing#scrollTo=2F54yHQPpBcT)
    - [POI classification](https://colab.research.google.com/drive/13Nzf0HhnP8UnsrCYeunXlha72LFBLH1W#scrollTo=NmetwGSns3Za)
    - [Build customer demographic profile](https://colab.research.google.com/drive/1J1dvPRfiyQf6pjaRuLMpcvNF2f1wy3xN?usp=sharing)
  + Safegraph’s internal community “help”
  + Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic (Kang 2021)

[DELPHI COVID-19 Trends and Impact Survey (CTIS):](https://delphi.cmu.edu/covid19/ctis/)

* Carol hasn’t worked directly with dataset but Emory has a datause agreement with Ben/Kristin having access
* [Variables that may be of interest:](https://delphi.cmu.edu/covidcast/survey-results/)
  + Does the oldest child in your household (pre-K–grade 12) go to in-person classes full-time?
  + In the past 24 hours, did you go to an indoor market, grocery store, or pharmacy?
  + In the past 24 hours, did you have a meal or drink indoors at a bar, restaurant, or cafe?
  + In the past 24 hours, did you use public transit?
* Relevant literature:
  + [The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination](https://www.pnas.org/doi/10.1073/pnas.2111454118)
  + List of publications at end of this link: <https://delphi.cmu.edu/covid19/ctis/>

**To-dos for Erica:**

* Review literature
  + What has already been done/explored with the two datasets?
    - Not much has been done with the DELPHI data set and I have definitely not found anything pertaining to both datasets/comparing or using data from both datasets
    - There is a lot of research from 2020-2022 with the Safegraph data looking at mobility trends
    - I will compile a list of research articles with a summary of the information presented for our next meeting
      * This article “[Association of Mobile Phone Location Data Indications of Travel and Stay-at-Home Mandates With COVID-19 Infection Rates in the US](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7489834/)”
        + This is a great article that looked at SafeGraph data and compared to trends in COVID-19 transmission

They used the amount of time at the “home” location to see if the stay at home mandates worked

* + - **Laying out an introduction with these articles**
      * Use bullet points
      * Why it is important
      * Where have these datasets been used
      * Lay it out as paragraphs for an intro
      * DELIVERABLE --------!!!!!
  + Why is important to groundtruth mobility data from app-based geolocation?
    - To show that the data is relevant
    - How would we be able to ground truth both data sets?
      * Would it be best to use the facebook data as conformation of the trends we are seeing in the other data?
  + **Harvard seminar about groundtruthing mobility data --- presentation**
    - <https://www.crisisready.io/events/ground-truth-matters-travel-surveys-covid-19-and-the-limits-of-human-mobility-data/>
* Carol is going to share code
* Explore the two datasets
  + Try starting with one week of Safegraph data and creating simple tables
    - There is soooooo much data….it is hard to find a place to start
    - When we are looking at this data set are we going to be looking at raw visit counts or normalized\_visits\_by\_total\_visitors?
      * How was the data normalized?
        + **They count the number of devices in each of the census block group using nighttime data over 6 weeks to say this is a home**
    - How do you recommend starting to look at this data
      * I can produce tables with counts by place but I am not sure if that is going to help us
      * Is there a way to look at the movement of people with this data and if so how would this be best to look at?
      * **Look at time trends!** 
        + For loop to look at number of trips to specific location by the hour
        + Save result in a list to summarize results

Count the number of rows

Do.dot.call rbind to combind into one diplir function

* + - How was this data processed? From what I saw online it looks like it comes in CSV’s??
  + **INTERNAL linking of place key to location FROM safegraph**
  + Some fields from Safegraph are “JSON” fields, there are a few functions in the [SafeGraphR](https://safegraphinc.github.io/SafeGraphR/) package that are needed to “explode” these fields
    - Try “exploding” a few fields and create tables/graphs to explore them to understand how this works
    - Just started to explore the this package. I definitely need more practice with looking at this type of data.
    - -county -> cencus track -> cenus block (500- 2,000 people per group)

**Things to think about:**

* Are there other data sources that could be used to “groundtruth” the mobility data?
  + We could use census data as well as social distancing guidelines for the counties we are looking at to help groundtruth what we are seeing if needed
* What scale and time period are you interested in?
  + I think that depends on what we are trying to get out of the data. I think we might see less correlation when looking at time with stricter times of social distancing vs higher correlation when there is a release of social distancing
  + **Difference between where people are located ; high correlation in the begining**
  + **OVER 2 year period**
  + **Also look at voting county -> special regression**
  + **Outcome one correlation per- county ()**
  + **Increases and decreases for shopping in different measures ()**
    - **3 different outcomes**
* What geographical unit of analysis should both datasets be summarized to for comparison?
  + County, census tract of census block group?
  + [For DELPHI, aggregated data available for API download at the county-level](https://cmu-delphi.github.io/delphi-epidata/api/covidcast.html), if we want lower granularity, need to contact them
  + I think county level would be best; from what I have read up on SafeGraph the lower granularities the move varying the data
  + Looking at the variation between these options in our dataset -> county might be best
* Any other questions that use the two datasets that are of interest to you?
  + (I’m open to other ideas)
  + No I think this question would be good; though it might be interesting looking at the correlation between these data sets at differing levels of social distancing guidelines or differing levels of COVID-19 cases
    - With the thought being that the stricter the guidelines the less correlation we would be seeing; this might also be true for the levels of COVID-19 cases (as case rise we might see less correlation between the data sets)

Proposed Thesis Timeframe:

Find out about external deadlines

* Complete the thesis proposed outline
  + By next Monday
    - ALSO create a time dependent series from the data
      * Do this for Safegraph by type of location NIAX code
        + MIGHT HAVE the best spot to join
        + Need to join the NIAX data(point data) to the data in the folder
        + How to logically make what the safegraph data match to the DELPHI data

Make decisions around what would be best

Go through this and groups based on what I need – then group with POI then -> merge the sets

* + - What other co-varients we want to look at
      * Other predictors that might belong in this association
      * Do this for the DELPHI
* Main lit review is done
  + Need to find additional papers for method portion for evlaluation of mobility and survey data
* Start with county level then move down as need
* Download the data set from DELPHI
  + Add to one drive
* Katie – tough base with the methods section of the proposal
  + Special regression
  + How do you include time in that picture -> maybe timeseries
  + Seafegraph – out of all the number of trips what proportion visited the grocery store OR proportion of each county visited the grocery store on that day
* Send out timeframe before next meeting.