Predictive value of cell-phone geolocated mobility data collected from Safegraph on COVID-19 deaths.

**Introduction**

The COVID-19 pandemic has changed the way we look at infectious diseases. As the world has gotten more interconnected so has the ability of infectious diseases to transverse the world a pace that makes containment close to impossible. One of the mitigating factors in the spread and the speed of spread of infectious disease is human movement. When the predictions of infectious diseases effect on locations is looked at in an epidemiological light, most of the time this is through modeling the effect in SIR models rather than using the actual movement of individuals through a location. This changed as the data available around geolocated movements has become more available and necessary to accurately predict the effect of movement on disease spread.

Through the COVID-19 pandemic mobility data was used to assess compliance with social distancing and other recommended restrictions; measures of interaction trends were then used to understand potential spreading/super-spreading events resulting in predictions of outbreaks(Klise et al., 2021; Loo et al., 2021; Ye et al., 2021; Zachreson et al., 2021). This mobility data predictions allowed policymakers to efficiently set restrictions for the populace without the need to understand the specific disease dynamics and not have to take the time/money/effort in clinical trials(Ilin et al., 2021). The use of mobility data has been established to predict the location of an outbreak with quick and high accuracy, allowing the public to prevent further spread of disease(Schlosser & Brockmann, 2021).

Looking at the information present in the mobility data from Safegraph and grouping the data

provide a better understanding of the human aspect to mobile phone mobility data.. It is also important to realize the constraints of the collected mobility data by Safegraph, since to be a part of this data group you would need a phone and have downloaded specific apps which would ping your location give way to sampling bias (Coston et al., 2021).

**Methods**

Data was collected from the CDC, Georgia state government, the US Census, and Safegraph for information on GA counties between March 2020 and March 2022 to use in this model. The variables that were collected from the Safegraph data was movement to locations of interest, which were then standardized to state level scaling. To simplify the effect of movement to locations for this model, locations were broken up into four categories (transportation, grocery stores, restaurants and eateries, and k-12 education locations) by NIAC codes. Once the locations were identified to belonging to one of these four categories the average movement per week per county was calculated. The data collected from the CDC and Georgia state government were the daily deaths and vaccinations for COVID-19. This data was then standardized to county populations and summed to get weekly totals. The US Census provided data on race percentages, income level, age distributions, household size, and urbanicity. This data was taken from the latest available at the time of collection. The percentage of individuals who voted for Donald Trump in the 2020 presidential elections were also taken from the Georgia state government information online.

Once the data was collected and cleaned it was evaluated for correlation and collinearity before being split into training and test datasets at a 80/20 split. The data in the training set was then normalized to better evaluate the effects of the variables without the interference of differing scales. A multilinear regression was run on the training set and evaluated with the testing set providing information on the resisduals, F-value, and R2 of the model.