Predictive Value of Cellphone Geolocated Mobility, Vaccination, and Social Factors on COVID-19 Mortality to Provide Foundational Framework for Predicting Outcomes of Future Pandemics

Erica Lynn Johnson, Carol Liu, PhD Candidate, and Ben Lopman, PhD

Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA 30322

Abstract

Relevance

Current research into the factors associated with COVID-19 mortality have shown that social distancing has a direct effect on the levels of COVID-19 deaths[1]. To expand on this research, this study aims to find the predictive value in using cellphone geolocated mobility, vaccination, and social factors on COVID-19 mortality. Knowing that the effect of these variables are most likely not linearly associated to mortality, using a predictive model that allows for nonlinear relationships and is able to handle missing data and outliers will increase the predictability of the model.

Variables

This study assessed COVID-19 mortality as the main outcome. Mean movement aggregated to four categories (K-12 grade schools, food service locations, points of public transportation, and visits to grocery stores) in each county for each week were considered the exposure of interest in our models. We then added in covariates of vaccination, population density, GDP level, level of urbanicity, household size, age, and political affiliation to address cofounding effects of human movement on COVID-19 mortality.

Design

Data was gathered from all 159 counties in Georgia for dates ranging from March 2020 and March 2022 using SafeGraph, the CDC, GA state databases, and US Census data. After processing this data was visualized using correlation graphs, histograms, and scatter plots to check for collinearity and possible associations between variables. These variables were then evaluated using both simple and expanded linear and Gradient Boosted Trees (GBT) models. Model statistics were looked at to assess the model's performance and predictability.

Main Findings

Multiple models, looking at different ways to evaluate movement to locations within the categories were evaluated to provide the best way to include this data in future disease predictive models. We found that there was little difference between the three ways we looked at geo-located data and their effect on the model's ability to accurately predict COVID-19 deaths. The GBT models significantly out preformed the linear regression models. Expanded GBT models, which considered all covariates and exposures showed a good R^2 value around 0.6 with low MAE and RMAE values showing the high precision of this model. Though I believe that adding in mobility metrics into machine learning algorithms can be a useful tool for predicting pandemic outcomes in the future, finding the right way to estimate this metric is a problem that will need to be tackled, since grouping by categories and looking at the mean movement in these categories does not seem the best way to do this.

Introduction

- There has been approximately 6.9 million deaths worldwide, of which the US accounts for 1.1 million as of April 2023
- COVID-19 has disproportionately affected low-income communities in the US
- Burden of this pandemic can be seen at the individual level with impacts directly on health, economic stress, job and future uncertainty, and mental health impacts of isolation and other control measures
- There are currently multiple mRNA vaccines approved for COVID-19, which have shown high efficiency and efficacy
- Many factors have played a role in the spread of COVID-19 including but not limited top population density, socio-economic factors, politics, urbanicity, vaccination level, and age
- Mitigating measures of COVID-19 have included social intervention, testing and contact tracing, as well as pharmaceutical intervention
- Mobility metrics have been used to approximate human movement to understand the effect of social interventions as well as approximate transmission factors
- Machine learning in public health has been used in disease predictions, giving an advantage of being able to handle highly complex and nonlinear relationships between exposure and covariates with the outcome of interest, as Gradient Boosted Trees (GBT) model does
- The main objective of this study was to evaluate movement to specific locations (exposures) over time and covariates effects on the predictability of COVID-19 mortality

Methodology

Exploration Framework & Context

- Complete a literate review on current COVID-19 research and Machine Learning research in Public Health
- Locate data sources to collect both outcome and covariate information from
- Obtain this data from websites associated directly with or collect the data directly from the CDC, Georgia state government, the US Census, and SafeGraph

Data Processing

- Standardize SafeGraph data to represent visits by state level to each POI
- NIAC'S codes were used to label SafeGraph POI data points to be in one of four specific categories: K-12 schools, food service locations, grocery stores, & public transportation locations
- Labeled POI locations were aggregated by county then week and summarized to obtain the mean movement in each category for each county each week
- Data for COVID-19 deaths & vaccinations rates were aggregated to weekly values
- These rates were then converted to be amount per 1,000 individuals by county population numbers
- Other covariates representing county specific values were standardized by overall county population

Covariate Selection & Data Visualization

- Data was graphed in scatter plots, histograms, and correlation plots to evaluate for collinearity and associations
- Changes in COVID-19 deaths and movement overtime were graphed on animated maps
- Features were selected based on literature and graphical results
- Three data sets were created looking at movement data as: 1. Continuous variable, 2. Dichotomous variable, and 3. A three level categorical variable
- A lag graph was used to analysis the effects of time on these movement categories

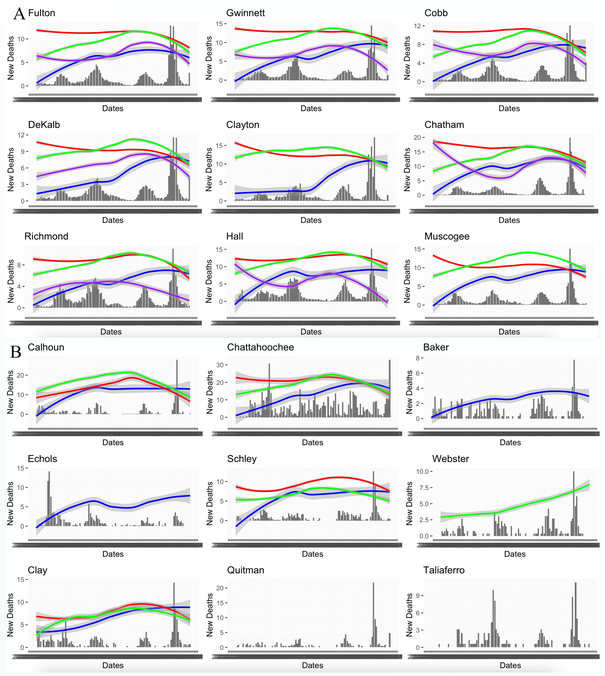
Model Selection & Training

- Data was split into an 80:20 ratio of Training to Testing data by randomizing weekly data regardless of geo or time location
- Initial linear regression models were run on data using dates and locations as included variables
- Data was changed into timeseries format and bundled into a list based on county
- Resulting list was split into a 80:20 ratio of Training to Testing data by randomizing counties, keeping the timeseries together
- GBT model was created and trained on training data before using testing data to evaluate predictability

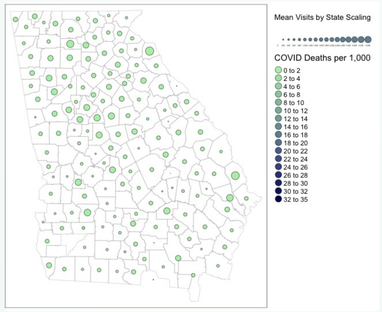
Model Performance Evaluation

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Coefficient of determination (R-squared), but adjusted and not adjusted
- F-Statistic
- Parameter estimates

Results



Histogram of COVID-19 deaths per 1,000 individuals in the (A) top 9 and (B) bottom 9 Georgia Counties effected by this pandemic. Lines of movement in the county was added to each graph, scaled and smoothed to be seen as the overall effect over time with the following representation of movement categories: red – education, blue – grocery, green – food services, and purple – transportation



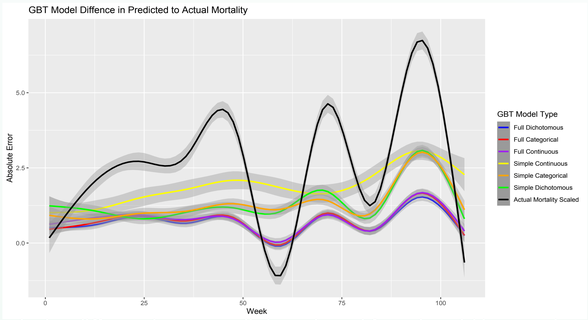
Frame captured from animated maps showing movement in the grocery category in the first week and the association with COVID-19 mortality

Linear Model	Multiple R-squared value	Adjusted R-squared value2	F-Statistic	K-12 Schools Estimate	Grocery Stores Estimate	Food Service Estimate
Simple Linear Model with Continuous Movement Variable	0.01265	0.01246	67.45	0.00011*	-0.00004*	-0.0001*
Simple Linear Model with Dichotomous Movement Variable	0.004917	0.004728	25.91	0.3017*	0.04115	-0.0632
Simple Linear Model with Categorical Movement Variable	0.0205	0.02013	55.08	Medium: 0.56723* High: 0.75827*	Medium: -0.06754 High: -0.02113	Medium: 0.01586 High: -0.18447*
Expanded Linear Model with Continuous Movement Variable	0.07103	0.06975	55.35	8.751*	1.614*	-4.119
Expanded Linear Model with Dichotomous Movement Variable	0.06873	0.06744	53.21	0.3478*	0.1610	0.02507*
Expanded Linear Model with Categorical Movement Variable	0.06935	0.06799	43.72	Medium: 8.182* High: 10.902*	Medium: 0.305 High: 2.823*	Medium: -0.916 High: -4.022*

*Linear model evaluations, with *denoting significances at an alpha of at least 0.001*

GBT Model	R-Squared Value	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
Simple GBT with Continuous Movement Variable	0.0004004477	1.891814	3.379743
Simple GBT with Dichotomous Movement Variable	0.004000327	1.34146	2.006809
Simple GBT with Categorical Movement Variable	0.02189969	1.328132	1.979287
GBT with Continuous Movement Variable	0.6067073	0.7437192	1.3446
GBT with Dichotomous Movement Variable	0.6173938	0.6966837	1.271944
GBT with Categorical Movement Variable	0.6042525	0.7437192	1.313551

GBT Model performance of different movement variable data types



Absolute difference between predictive and actual for each model in comparison the scaled GA overall mortality per week

Conclusion & Public Health Implications

- From the histograms represent the epidemic curve of the counties COVID-19 deaths per 1,000 individuals, with smoothed lines of movements within the four categories scaled to show overall trajectory, we can see overall in these graphs a trend of four distant peaks of COVID-19 deaths, which are likely correlated with the waves of COVID-19 as they moved through the GA counties
- From the animated maps, movement to K-12 grade schools showed some correlation between increase in movement to these locations and higher COVID-19 mortality
- For the linear models, the multiple and adjusted R square values are very close to zero and the F-statistics are high, resulting in competing interpretation of the model results. This most likely indicates that the model has statistically significant predictors, like our exposure variables, but that they collectively, even in our expanded linear models, only explain a small proportion of the variation in COVID-19 mortality.
- The coefficients values of our exposures do show a significant impact in our linear models, with movement to K-12 school estimates being significant in every model with a positive increase association
- All GBT expanded models significantly out preformed all (extended and simple) linear models, resulting in the implications that the covariates account for some of the cofounding effects of movement on COVID-19 mortality and that the association between movement as well the covariates and COVID-19 mortality is most likely a nonlinear
- From the model evaluation it seems that the way we modeled human movement, regardless of the type of variable this we put into, does not have a direct relationship to COVID-19 mortality when looking at only the simple models