Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Erica Lynn Johnson Date

Predictive value of cellphone geolocated mobility, vaccination, and social factors on COVID-19 mortality to provide foundational framework for predicting outcomes of future pandemics

By

Erica Lynn Johnson
Masters of Public Health

Epidemiology

_____

Ben Lopman, PhD
Committee Chair

_____

Carol Liu, PhD Candidate
Committee Member

Predictive value of cellphone geolocated mobility, vaccination, and social factors on COVID-19 mortality to provide foundational framework for predicting outcomes of future pandemics

By

Erica Lynn Johnson

BS, Mathematics
The Pennsylvania state university
2020

BS, Microbiology
The Pennsylvania state university
2018

Thesis Committee Chair: Ben Lopman, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2023

Abstract

Predictive value of cellphone geolocated mobility, vaccination, and social factors on COVID-19 mortality to provide foundational framework for predicting outcomes of future pandemics

By Erica Lynn Johnson

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 (visits to K-12 grade schools, visits to 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 models 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.

Predictive value of cellphone geolocated mobility, vaccination, and social factors on COVID-19 mortality to provide foundational framework for predicting outcomes of future pandemics

By

Erica Lynn Johnson

BS, Mathematics
The Pennsylvania state university
2020

BS, Microbiology
The Pennsylvania state university
2018

Thesis Committee Chair: Ben Lopman, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2023

Acknowledgements

I could not have completed this project without the help and guidance provided by both PhD candidate Carol Liu and my thesis advisor professor Ben Lopman, PhD. It is through the mentorship and guidance of these individuals that I was able to have complete this project. I would especially like to thank Carol, who help to keep me on track and worked through all the problems a project with a lot of data can run into. It was with her help and guidance that I was able to apply my interest in machine learning to this epidemiological data.

Introduction

Burden of Disease

As of April 2023, there have been over 760 million reported cases of COVID-19 worldwide and approximately 6.9 million deaths attributed to COVID-19. The United States has one of the highest overall cases (102 million) and deaths (1.1million) ascribed to COVID-19, making research into prevention of COVID-19 a virial topic, as our numbers have accounted for about 1/6 of the total number of cases and deaths worldwide[7–9].

COVID-19 has disproportionately affected low-income communities in the US, showing how the intersection of poverty, economic instability, and systemic racism has led to dire health consequences. This has had the effect of further intensifying the disparities that existed in the US prior to COVID-19, exacerbating the disparities in health equity and deepening the divide between communities[10].

The burden of this pandemic can be seen in more than just the effect it has had on individuals and the health system, as part of the control measures for containing the spread involved shutting down non-essential jobs greatly impacting the economy. These measures, though important to preventing further spread of COVID-19, have been found to have significant mental health consequences due to increased rates of anxiety, depression, and other mental health conditions, as well as substance abuse and domestic violence during this time of social isolation, economic stress, and uncertainty about the future[11,12].

Natural History of COVID-19

SARS-CoV-2, the causative agent of the COV1D-19 pandemic is a coronavirus, and one of three notable coronaviruses that have led to outbreaks in the human population[13]. Initial reported cases were from the Wuhan City, Hubei Province in China, who reported to the WHO office in December 2018 of unexpected increases of individuals with pneumonia of unknown cause. This prompted an investigation where the novel cases were identified as SARS-CoV-2. A pandemic was officially announced by the WHO on March 11, 2020, with cases popping up across Europe and Asia[14]. 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 at a pace that makes containment close to impossible.

The presentation of COVID-19 in each individual infected with the virus has been diverse, with no one symptom being reported in all cases. The most common symptoms are cough, fever, myalgia, chills, fatigue, headache, and shortness of breath[15]. Individual infected with COVID-19 could also present as asymptomatic, which is the case for about 40-45% of individuals[16]. Symptoms can start anywhere from 2-14 days at with the infectious period lasting up to 20 days (or longer depending on severity of symptoms)[17].

Current treatment for COVID-19 includes three currently emergency approved antivirals, one of which requires continued infusion over three days at a health care facility. Convalescent Plasma treatment has also been approved for treatment of COVID-19.

Known Covariates of Infectious Diseases

As with any infectious disease, there are many factors that affect the ability of the disease to spread, and this is no different for the SARS-CoV-2 virus. In many of the currently published articles evaluating both the spread of the virus and effective control measures we find that not only does geography and timing play a role but also factors like population density, socio-economic factors, politics, urbanicity, age, and many other individual and community attributes play an important role[18–23].

When looking into what causes a respiratory infectious disease to spread one of the main concerns is the amount of individual within an area. To account for this in a model, we can look at things like population, urbanicity, and household size which give an overall picture of the community in a specific location. Adding in social and other non-measurable factors, like the level of discrimination that minorities face, are slightly harder to account for. It is know that in the US that race plays a role in health outcomes and access to health care, not because it is a factor in the actual disease but due to inequitable access to treatments, housing locations, and other factors responsible for social inadequacies in the US. Using variables like GDP, household income, and percentage of Black identifying individuals can help to account for these factors in modeling infectious diseases. With COVID-19 specifically it has been shown that age has a direct effect on the outcome of infection, so accounting for the population in a county over the age of 65 can help even out the model to make it more generalizable.

Mitigating Measures of the COVID-19 Pandemic

As COVID-19 spread and the pandemic was declared various measures have been implemented across the world to mitigate its spread. These mitigation measures included social interventions such as social distancing and the use of face masks. Campaigns were put forward for hand hygiene, and any locations were people gathered, like grocery stores and churches, were encouraged to frequently wash down any surfaces with high constant rates like door handles.

Other mitigation measures include testing and contact tracing to identify individuals who have potentially been exposed to COVID-19 and quarantine and isolation measures to prevent further spread. All of these measures have been shown to be effective in reducing the transmission of the virus and limiting the impact of the pandemic on public health. In addition to these social measures, pharmaceutical interventions such as vaccines have also played a critical role in mitigating the impact of the COVID-19 pandemic.

COVID-19 Vaccination

With the emergency use approval of mRNA vaccines from Pfizer-BioNTech and Moderna having over 90% prevention rates, vaccines are an important time varying mitigation measure.

In addition to these vaccines being highly effective in reducing the transmission of COVID-19 they have also been shown to prevent severe illness and hospitalization. However, vaccine hesitancy remains a significant challenge in some communities, and there are concerns regarding the equitable distribution of vaccines globally. One of these factors that has shown to influence vaccine hesitancy is political affiliation.

Mobility & SafeGraph

During COVID-19, mobility metrics that approximated human movement were used to understand adherence to social distancing and subsequent relaxation of behavior after the most stringent protocols were rolled back. Mobility data also served as a proxy for changes in social contact in transmission models where it provided a direct, quantifiable link between observed behavioral changes and potential changes in population-level transmission. Increases in human movement observed by mobility data were used as an early indication of potential rises in cases and were tracked by public health agencies such as the Centers for Disease Control (CDC) and Georgia Department of Public Health (GDPH). Our analysis seeks to explore the predictive value of cell-phone geolocated mobility on COVID-19 deaths to determine its utility as an early indicator for upcoming waves during an epidemic.

Through the COVID-19 pandemic, mobility data was used to assess compliance with social distancing and other recommended restrictions. Measures of mobility trends were then used to understand potential spreading or super-spreading events resulting in predictions of outbreaks[24–27]. Mobility data described the connectivity within and between communities, allowing for predictions of the spatial patterns of disease spread with quick and high accuracy. This allowed policymakers to efficiently set restrictions for the populace prior to detailed knowledge of specific disease dynamics, and effectively preventing further disease spread[28].

Machine Learning in Epidemiology

Regression models provide an understanding of the association between exposure and outcome of interest and machine learning algorithms can make use of regression models to provide a predictive quality to the results. Unlike general regression models, machine learning algorithms are able to capture complex patterns and relationships in data without relying on explicit assumptions or pre-defined models. They are able to handle highly complex and nonlinear relationships between predictor variables (exposure and covariates) and the outcome of interest which makes them a powerful tool in prediction tasks[29].

Machine learning algorithms are not a new topic in public health and have been used by epidemiologist and statisticians before in disease predictions. But the pandemic spurned a need for the most accurate predictions of disease spread, prompting growth in the popularity of using machine learning in public health and infectious diseases. These models have shown great promise in predicting COVID-19 outcomes using a range of data sources, including demographic and socio-economic data, health indicators, and mobility patterns. Supervised machine learning algorithms allow the researcher to use regression models on labeled data to train then evaluate the models' ability to predict a specified outcome[26,30–33]. Several studies have explored the application of machine learning models to predict COVID-19 spread, hot-spots, morbidity, case numbers, and mortality. These models included the use of random forest[34], ARIMA models for time-series forecasting[34], Gaussian process regression[35], linear discriminate analysis[36], suggesting many feasible approaches to achieve our research goals.

Gradient Boosted Trees (GBT) models have several advantages over other machine learning algorithms. GBT models are less prone to overfitting, a problem that happens when the model is too complex and forms to the training data without generalizability to un-seen circumstances. They allow for the input of many features regardless of variable type (categorical/continuous), as well as allowing the regression to capture non-linear relationships between these features and COVID-19 mortality. GBT differs from traditional regression models in its ability to handle complex and nonlinear relationships, its built-in mechanisms to handle overfitting, and its effectiveness in handling missing data and outliers[37].

We use the GBT model to see if the effects of human movement coupled with additional covariates provide an accurate prediction of COVID-19 deaths at the county-level and over weeks of the pandemic [38].

Impact and Objectives

Predicting the spread of COVID-19 and identifying regions that are most vulnerable to severe outcomes, such as deaths, is critical for public health officials to make informed decisions about resource allocation, planning and policy decisions. But this does not only apply to COVID-19. Through modeling the effect of movement on COVID-19 mortality, we will be able to provide future respiratory outbreaks a starting point to predict the effect of movement on that novel disease. The overall objective of this study was to look at the predictive value of cellphone geolocated mobility, vaccination, and social factors on COVID-19 mortality. To tackle this objective a GBT model was designed, trained, and tested for this study.

Methods

To analyze the relationship between human movement and other covariates such as vaccination rate, population density, GDP level, level of urbanicity, household size, age, and political affiliation on COVID-19 mortality rates in GA counties from March 2020 to March 2022 the data was initially collected and processed to scale based on counties population. Further processing was done the main exposure variables to look at model changes when these variables are coded as continuous, dichotomous, and categorical. The lag between these variables and COVID-19 mortality rates were graphed to observe any effect and a two week lag period was applied to the vaccination data to appropriately model the initiation of protection given by these vaccines at 14 days after the second dose[39].

After processing the data, both a linear and non-linear GBT machine learning model were used to analyze the effect of the exposure and covariates on our outcome and test our hypothesis that human movement had a direct effect on the predictability of COVID-19 mortality. The flow of this overall process can be seen in Figure 1.

Exploration Framework and Context

The study at hand was conducted using a retrospective framework with county-level COVID-19 mortality data collected from March 2020 to March 2022 in the state of Georgia, USA.. One of the most versatile and accurate machine learning methods, GBT model, was used to carry out this analysis. This process involved five main components: Data Collection, Data Processing, Feature Selection and Data Visualization, Model Selection and Training, and Evaluation. Figure 1 shows the steps taken during this process. This study used only R programming language to perform all steps of this process.

Data Source & Processing

In this research study we have taken geolocated time series data from Georgia counties between March 2020 and March 2022 to analyze SARS-CoV-2 deaths. To better understand the effect of human movement on the COVID-19 pandemic we decided to model the outcome of COVID-19 deaths. Though there was availability on the number of cases of COVID-19 in each county each week, the case number is known to be under-reported especially as time progressed and at-home test became more readily available. To model the effect of movement on COVID-19 mortality, we considered our exposure to be current average movement in four categories: to educational institutes k-12, to grocery stores and markets, to food service locations, and transportation (both busses and plane locations).

Outcome

To measure the effect of human movement in a way that was comparable between counties of different population levels, mortality data from the CDC was collected and converted into the number of COVID-19 associated deaths per 1,000 individuals living within that specific county. Before this was done, to allow for a timeseries analysis, the data was aggregated into weekly counts, based off of the weekly dates from SafeGraph *(Table 1)*.

Exposure

We used data gathered from SafeGraph as a proxy for human movement. The mobile phone-based geolocation data offered by SafeGraph is obtained from anonymous sources across the United States. This data is sourced from mobile phone users who have enabled geolocation on undisclosed

mobile applications. SafeGraph assigns the devices represented in the data to their respective home census block group based on the most frequent nighttime location during the preceding six weeks. SafeGraph has designed a privacy-centric approach to handle the data by pre-aggregating it into weekly visit and visitor counts to Points of Interest (POIs). POIs, which are public locations such as stores, schools, parks, health facilities, offices, hotels, among others, are bounded by polygons, and devices are counted as having visited a POI when a ping is captured within the polygon for at least four minutes. Additionally, the data is further categorized into weekly Census Block Groups (CBG)-POI visitor flows to provide a more granular level of analysis.

When receiving the data from SafeGraph, specifically it's Places dataset, rows are organized into POI's and the data for each of the POI's standardized at the state level for visits to each location reported on. POI's were then categorized into four different categories of interest: visits to K-12 grade schools, visits to restaurant and other food service locations, points of public transportation (including airports and bus terminals), and visits to grocery stores. This was done using all applicable location with a NIAC's code in one of the four categories to label the data. The these labeled categories were grouped by county location then by week. This was done to get a single mean value for each category in each county per week as a way to approximate the movement in a specific county to these four location types each week.

We hypothesized that the association between changes in movement and COVID-19 mortality may not be linear and decided to look at the effect of grouping movement in different ways on the models' predictability. This was done by first using the continuous mean movement value to the four locations. We then dichotomized this value based on sectioning as the following: movement

less than the overall movement average for that category was given a value of zero, and movement above the overall movement average for that category was given a value of 1. Lastly, a categorical approach was also looked at with overall movement in each category broken into low, medium, and high based off of one-third and two-third break points. Table 2 shows the three ways SafeGraph movement data was looked at for all three categories in our separate model's analysis. In addition, since it was likely that the movement in one week would not directly impact that week's COVID-19 deaths, rather the effect would lag behind, a graph of lag effects on the GBT model were done for the different movement groupings from zero to 14 weeks as the potential lag component *(Figure 2)*.

Other Covariates

In consideration of the currently published research covariates such as county population density, GDP level, level of urbanicity, household size, age, political affiliation, and vaccination coverage were added to the model to improve its predictive capability and generalizability[18–23,40–43]. The data on these different confounders was gathered from many sources across the web, but in totality they can all be traced back to four main sources: the CDC, Georgia state government, the US Census, and SafeGraph as shown in Table 4. In addition to this information, website references can be found in Appendix 1 for exact locations of this data online.

After collecting the above variables from multiple sources, pre-processing of the dataset occurred. This involved aggregating data to county weekly values, rates of vaccinations per 1000 individuals living in the county, lagging the vaccinations for 2 weeks, and handling missing or incorrect values.

Covariate Selection and Data Visualization

Understanding the relationships between the variables to be used in this analysis was an important first step in the process of coming up with the correct model. To do this a comparative graph which places each variable against all other variables was done. This graph includes only time independent variables as well as total COVID-19 deaths per 1000 individuals for each county. This comparative graph included both histograms and scatter plots. In addition, a correlation matrixes was created between each of the time independent variables, and totals vaccinations and deaths for all time.

This was done to visualize any trends, like a linear association, and discover any possible collinearity that would need to be addressed before moving forward in this process. Exploration statistics such as total values, median, mean, standard deviation, range, and more were done on covariate variables to give an understanding of distribution and counts. As stated in the descriptions of a few features in Table 2 and Table 3, three variables (normalized visits by state scaling to transportation terminals, counties total votes in 2020 presidential election, and median age of county from 2019) were removed after pre-processing the data and not used in our models.

The SafeGraph transportation category was not used in this analysis due to there being less than one third of the counties with transportation information. Both counties total votes in 2020 presidential election and median age of county from 2019 were not used due to collinearity issues with other variables.

Model Selection and Training

To tackle the effect of human movement on COVID-19 mortality, a GBT model was designed, trained, and tested. The GBT model is a machine learning algorithm used for both regression and classification problems and was used here for our regression analysis. GBT is a decision tree-based ensemble learning method that combines multiple decision trees to make more accurate predictions. The GBT model uses an iterative approach after it has been initialized. Starting with a simple model to predict the average value of COVID-19 mortality, the GBT model then goes on to build a decision tree to predict mortality based on the features included in the model. It makes what is called a weight adjustment, changing to value of observations based on their residuals (differences between the observed values and the predicted values of COVID-19 mortality). Additional trees are added to the model based on the residuals from the previous iteration, improving its predictions by focusing on the observations that were poorly predicted by the previous trees. This repeated process happens for a specified number of steps with the final prediction being the sum of the predictions of all the trees. GBT models have emerged as a powerful machine learning tool that can make accurate predictions on complex datasets [44–46].

After initial processing and feature selection a simple linear regression model was run using only the three movement variables, which included visits to K-12 grade schools, to restaurant and other food service locations, and to grocery stores, as features to look at the effect on our outcome, COVID-19 mortality. After this a linear model containing both exposure variable (movement to the labeled categories) and identified covariates (vaccination rate, population density, GDP level, level of urbanicity, household size, age, and political affiliation) was created and processed as a multi-linear regression model, which used each time step (week) as a single regression.

Before applying any machine learning algorithms, our data was converted to a timeseries list, containing one object per county. This list was then broken with a 80:20 ratio using randomization on the counties, keeping the timeseries together to improve accuracy of our results. For each training and test set the data was then broken into outcome (new deaths per 1000 individuals) and predictor variables. A GBT model was applied with 50 rounds using xgboost package in R, and trained using the training data set. A prediction on the COVID-19 mortality in each of the testing counties each week was then made using the testing predictor values with this trained model. This process was repeated three times, with each movement variable type. We did this approach to see if coding the exposure variable in these different ways would effect the predictability of the model overall, whit the assumption that using a continuous outcome (movement to each category in each week given as a mean over all labeled POIs within that county and category) would have a positive effect on the models predictive value.

Model Evaluations

For all models, both the linear regression and the GBT models, the coefficient of determination was evaluated. For the linear regression models F-statistic and adjusted R-squared were also calculated and parameter estimates for exposure values were evaluated. For the GBT models the predicted values were compared to the actual values to provide an evaluation on the model through calculating the mean absolute error (MAE) and root mean squared error (RMSE).

We looked at both the linear model and the GBT predictive model to observe potential differences in model output if we considered the variables effect on COVID-19 mortality to be non-linear.

Results:

Data Description & Visualization

Figures 3 and 4 show the top and bottom nine counties in Georgia, regarding overall deaths. These histograms represent the epidemic curve of the counties COVID-19 deaths per 1,000 individuals. Smoothed lines of movements within the four categories were scaled to show overall trajectory, allowing comparisons between the outcome and exposure. 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.

Figure 5 shows this comparative graph where is histogram is representative of the counts of that variable, and the scatter plots are pointwise effects of one variable on another. We were able to observe collinearity between age variables, household size and age variables, as well as percentage of the population who identify as Black and the percentage of the population who voted for Trump in the 2020 presidential elections.

To continue this simple analysis of the variables effect on each other a correlation heat map was created, Figure 6 and animated maps were also created to show the change of movement (in each category) and number of COVID-19 deaths per 1000 individuals in each county (links to these are in Appendix 2). An example of one frame of these maps is seen in Figure 7. These animated maps showed movement to K-12 grade schools showed some correlation between increase in movement to these locations and higher COVID-19 mortality.

Linear Models:

Linear models on both simple, including only movement as our exposure variables, and expanded, including all covariates, all had somewhat similar metric values as seen in Table 5. The R-squared values for these models was very close to zero for all of them but shows a worse result when we dichotomous the movement data as well as simple models having a worse result then expanded models. Though the resulting multiple and adjusted R square values for these models range from 0.004 to 0.069 (very close to zero) and the F-statistics are high, ranging from 25-67, 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.

We see the highest impact of these exposures in the expanded linear model with categorical movement variables. In that outcome model both movement to K-12 schools and grocery stores show a positive association for both medium and high movement coefficients of 0.305(medium) and 2.823(high) for grocery stores and 8.182(medium) and 10.902(high) for K-12 schools. Regarding the estimates for food service, there is a negative association between movement to this category with estimates of -0.916(medium) and -4.022(high). All high categories resulted in a significant coefficient estimate, and K-12 schools also had a significant estimate for medium factor value.

Though we see the most of the time at least two of these exposure variables have an impact on the model that is not just a result of random variation in the data, with low R-Squared values and high F-statistics we are getting a 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.

Gradient Boosted Trees (GBT) Model:

As with the linear models simple and expanded version of separate GBT models were run on our selected data. Table 6, shows the overall performance measurements used to evaluate the different GBT models, and Figure 7 is a smoothed graph of the difference in the models predictive value and the actual value of deaths per 1,000 individuals.

Before these models were run, a lag graph for all four movement categories was created and analyzed to see the effect of lagging the movement on the predictive value in each of the GBT models. These graphs look at the error in prediction through MSE and RMSE, and the goal of looking at these lagged values of error is to see which lag value produces the lowest error for each GBT model. In this case we notice an increase of error as we increase the lag. This shows the optimal lag for these variables is zero, so no lag was applied to the exposure variables (movement to the four categories of POIs) in both the linear and GBT models analyzed in this study (Figure 2).

As with the linear regression models in Table 5, there is not much change in the model statistical metrics when changing the exposure variable type when comparing the simple GBT models to each other or full models to each other. We do see that all the full GBT models completely out preform the simple ones as well as all of the linear regression models. When looking just at the R-squared values of these different models, the GBT model with the simple regression does just as

poorly as the linear model ranging from 0.0004-0.02, but once covariates are added in the GBT model performs significantly better than any linear model with R-squared value ranging from 0.60-0.61. With MAE values ranging from 0.69-0.74 for the expanded GBT models this suggest that on average the predicted COVID-19 deaths per 1,000 individuals from the models have an absolute difference of 0.69 to 0.74 units from the actual observed values. These lower MAE values indicate that the GBT model performance is good since they reflect smaller errors between predicted and actual values for COVID-19 deaths per 1,000 individuals. The RMSE values, which measure the average magnitude of error between the actual and predicted values of the model, range from 1.27-1.34 for the expanded GBT models. With these low RMSE values, it shows that these GBT models have performed well.

In tables 5 and 6 we can see that the full GBT models have R-squared values ranging from 0.6 to approximately 0.62, whereas the linear regression models and simple GBT models have R-squared values range from approximately ~0 to 0.07. This is a large difference in values, which results in a p-value of less than 0.0001 when doing a two sample, two tailed t-test on this data. These are significantly different values show that the expanded GBT models not only out preformed the linear regression models with an average R-squared of about 0.61 but that together the model data is able to explain over 60% of the variability in the mortality data.

Discussion:

Main Findings

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. Even

with this finding, from looking at R^2 values of our simple models (both linear and GBT), we do see that the full regression GBT models due show both high accuracy and precision, this indications that this model is accounting for and predicting some of factors associated with COVID-19 mortality. Knowing from previously published articles that there are many factors that come into effect of predicting COVID-19 mortality, this finding is not surprising. With an R^2 of around 0.6 for all GBT expanded models, we know that our exposure (travel to specific locations) and the chosen covariates have an interpretation value of 60% on the cause of COVID-19 mortality. This increase in predictability of GBT models that include all exposure and covariates could be due to correctly accounting for cofounding effects of our movement exposures on COVID-19 mortality.

In general, we see the trend of increased model performance when evaluating the full expanded model's vs any of the simple (exposure only) models. Knowing that the model's R^2 values, when only containing our movement exposure values is so low, it shows that the assumed relationship between movement and COVID-19 mortality was not adequately modeled here. This could be due to how we approximated human movement, by using aggregated data from only four specific categories. In our linear models we did see that the movement in the K-12 school category was significant in all models, regardless of how the variable was code in the model. In future analysis isolating this effect and looking more individually at travel to these location might prove to tighten up the predictability of the GBT model, rather then having these values aggregated per week in each county.

Limitations

As the GBT expanded models only account for 60% of the variability in COVID-19 mortality, we know that there is still 40% of variance in this outcome that is not covered by the association between labeled movement in our four categories and addition covariates on COVID-19 mortality. Adding in additional covariates, such as county health access, might provide a clearer outcome relationship.

In using a GBT model, we are also limiting our understanding of which factors have the most effect on the outcome predictability. It is not a simple process to evaluate the inner workings of a GBT model, and this was not attempted in this study, so evaluation of specific variables effects on COVID-19 mortality in the GBT model is a strong limitation to using this model in the future.

To continue to use this GBT model to predict COVID-19 mortality, or to use a similar model to predict a new novel disease's mortality would require continued updating to the training model. This would involve adding in new time series data, which in this model is movement to four location categories and vaccination rates. This is required due to the dynamic nature of the COVID-19 pandemic, which is something that is shared with any novel disease. Validation using rigorous scientific methods and ongoing monitoring of model performance is essential to ensure accurate and reliable predictions of COVID-19 mortality. In addition to using the proposed GBT models when looking at COVID-19 mortality, other sources of evidence and expert judgment to inform decision-making and policy development related to COVID-19 should be used as well.

Strengths

The GBT models out preforming the linear regression models is most likely due to the flexibility that is inherent to the GBT model type, that we cannot get when assuming a linear correlation. The GBT model, also allows us to add in our categorical variables with more effect, which we do not see in our linear models. In addition, movement, and the changes in movement that we are seeing in this data, even when looking at the animated maps, are not showing a simple association.

Though an R^2 value of 0.6 is good, a higher value was expected when using the GBT model to evaluate the predictability of movement and other covariates on COVID-19 deaths. It was initially thought that receiving a lower R^2 then what was expected might be due to the fact that neither of these models takes into consideration the lag response of social exposure to COVID-19 and the resulting death. But after producing lag graphs of the error in the different GBT models when lagging movement from zero to 14 weeks, the lowest error is seen when we do not add any lag to movement exposure variables in our model. This was an unexpected result, since on average after exposure to COVID-19 most deaths occur after four weeks.

The GBT models also have consistently low MAE values, showing that this model does have high precision in its prediction of the effect that movement and covariates have on COVID-19 mortality. As for the RMSE scores for the GBT model, these are also in the good range of RMSE scores allowing us to generalize these results to more than just Georgia.

Within the GBT models, there is not a lot of variation between the resulting model statistics and the change in movement variables. This could be due to the effect of movement still being captured the same, regardless of how we code this variable in the model.

Implications

Overall, this study highlights the effectiveness of using predictive models on understand effect of cellphone geolocated mobility, vaccination, and social factors on the predictability of COVID-19 mortality and warrants further study into this effect. These models can provide the groundwork into understanding novel respiratory viruses and help predict the outcomes of the next pandemic and it's spread across the word.

Public Health Implications and Future Directions:

The COVID-19 pandemic has caused significant public health challenges worldwide, with governments implementing a range of measures to control its spread. Social distancing and limiting travel outside of the home was one of these widely adopted measures used to curb the transmission of this virus. This study aimed to take into consideration confounding causes of not following social distancing measures, like political partisanship, overcrowding seen in large metropolitan areas, and racial disparities seen in our health care system.

The spread of any infectious disease is not limited to one person's response, but rather communities as a whole deciding to follow guidelines to keep from spreading or catching these diseases. The results of this study have important implications for public health policy, showing that political

partisanship and these other social factors are just as important in understand and controlling the spread of COVID-19 as social distancing guidelines.

In future studies, being able to look at other locations, and not aggregating location data into only four groups could benefit the model as well. Geo-located data is not the only way to measure human movement. Using survey response data, like what is available through Facebooks COVID-19 questionnaire, could provide a better account for the individual mindset, then cell-phone pinged data.

References:

1. Fowler, J. H., Hill, S. J., Levin, R. & Obradovich, N. Stay-at-home orders associate with subsequent decreases in COVID-19 cases and fatalities in the United States. *PLoS ONE* **16**, e0248849 (2021).
2. CDC. COVID Data Tracker. *Centers for Disease Control and Prevention* https://covid.cdc.gov/covid-data-tracker (2020).
3. COVID - Coronavirus Statistics - Worldometer. https://www.worldometers.info/coronavirus/.
4. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int.
5. News, A. B. C. COVID-19 has had 'devastating and disproportionate' impact on poorest Americans, report finds. *ABC News* https://abcnews.go.com/Health/covid-19-devastating-disproportionate-impact-poorest-americans-report/story?id=83893515.
6. Czeisler, M. É. *et al.* Mental Health, Substance Use, and Suicidal Ideation During the COVID-19 Pandemic — United States, June 24–30, 2020. *Morb. Mortal. Wkly. Rep.* **69**, 1049–1057 (2020).
7. Holmes, E. A. *et al.* Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *Lancet Psychiatry* **7**, 547–560 (2020).
8. Forchette, L., Sebastian, W. & Liu, T. A Comprehensive Review of COVID-19 Virology, Vaccines, Variants, and Therapeutics. *Curr. Med. Sci.* **41**, 1037–1051 (2021).
9. Machhi, J. *et al.* The Natural History, Pathobiology, and Clinical Manifestations of SARS-CoV-2 Infections. *J. Neuroimmune Pharmacol.* **15**, 359–386 (2020).
10. Burke, R. M. Symptom Profiles of a Convenience Sample of Patients with COVID-19 — United States, January–April 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, (2020).

11. Oran, D. P. & Topol, E. J. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Ann. Intern. Med.* **173**, 362–367 (2020).
12. CDC. Healthcare Workers. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html (2020).
13. Rocklöv, J. & Sjödin, H. High population densities catalyze the spread of COVID-19. *J. Travel Med.* taaa038 (2020) doi:10.1093/jtm/taaa038.
14. Dragano, N., Rupprecht, C. J., Dortmann, O., Scheider, M. & Wahrendorf, M. Higher risk of COVID-19 hospitalization for unemployed: an analysis of 1,298,416 health insured individuals in Germany. 2020.06.17.20133918 Preprint at https://doi.org/10.1101/2020.06.17.20133918 (2020).
15. Joy, M. & Vogel, R. K. Beyond Neoliberalism: A Policy Agenda for a Progressive City. *Urban Aff. Rev.* **57**, 1372–1409 (2021).
16. Matthew, R. A. & McDonald, B. Cities under Siege: Urban Planning and the Threat of Infectious Disease. *J. Am. Plann. Assoc.* **72**, 109–117 (2006).
17. Bi, Q. *et al.* Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919 (2020).
18. Grossman, G., Kim, S., Rexer, J. M. & Thirumurthy, H. Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 24144–24153 (2020).
19. Klise, K., Beyeler, W., Finley, P. & Makvandi, M. Analysis of mobility data to build contact networks for COVID-19. *PloS One* **16**, e0249726 (2021).

20. Loo, B. P. Y., Tsoi, K. H., Wong, P. P. Y. & Lai, P. C. Identification of superspreading environment under COVID-19 through human mobility data. *Sci. Rep.* **11**, 4699 (2021).
21. Ye, Y. *et al.* Predicting COVID-19 epidemiological trend by applying population mobility data in two-stage modeling. *Zhejiang Xue Xue Bao Yi Xue Ban J. Zhejiang Univ. Med. Sci.* **50**, 68–73 (2021).
22. Zachreson, C. *et al.* Risk mapping for COVID-19 outbreaks in Australia using mobility data. *J. R. Soc. Interface* **18**, 20200657 (2021).
23. Ilin, C. *et al.* Public mobility data enables COVID-19 forecasting and management at local and global scales. *Sci. Rep.* **11**, 13531 (2021).
24. Domingos, P. A Few Useful Things to Know About Machine Learning. *Commun. ACM* **55**, 78–87 (2012).
25. Ghouchan Nezhad Noor Nia, R., Jalali, M., Mail, M., Ivanisenko, Y. & Kübel, C. Machine Learning Approach to Community Detection in a High-Entropy Alloy Interaction Network. *ACS Omega* **7**, 12978–12992 (2022).
26. Firoozbakhtian, A. *et al.* Detection of COVID-19: A Smartphone-Based Machine-Learning-Assisted ECL Immunoassay Approach with the Ability of RT-PCR CT Value Prediction. *Anal. Chem.* **94**, 16361–16368 (2022).
27. Behnam, A. & Jahanmahin, R. A data analytics approach for COVID-19 spread and end prediction (with a case study in Iran). *Model. Earth Syst. Environ.* **8**, 579–589 (2022).
28. Ghafouri-Fard, S. *et al.* Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review. *Heliyon* **7**, e08143 (2021).

29. Painuli, D., Mishra, D., Bhardwaj, S. & Aggarwal, M. Forecast and prediction of COVID-19 using machine learning. *Data Sci. COVID-19* 381–397 (2021) doi:10.1016/B978-0-12-824536-1.00027-7.

30. Alali, Y., Harrou, F. & Sun, Y. A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. *Sci. Rep.* **12**, 2467 (2022).

31. Willette, A. A. *et al.* Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study. *Sci. Rep.* **12**, 7736 (2022).

32. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939785.

33. Youssef, A. *et al.* Development and validation of early warning score systems for COVID-19 patients. *Healthc. Technol. Lett.* **8**, 105–117 (2021).

34. Ferdinands, J. M. Waning 2-Dose and 3-Dose Effectiveness of mRNA Vaccines Against COVID-19–Associated Emergency Department and Urgent Care Encounters and Hospitalizations Among Adults During Periods of Delta and Omicron Variant Predominance — VISION Network, 10 States, August 2021–January 2022. *MMWR Morb. Mortal. Wkly. Rep.* **71**, (2022).

35. CDC. People with Certain Medical Conditions. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html (2022).

36. Feinhandler, I., Cilento, B., Beauvais, B., Harrop, J. & Fulton, L. Predictors of Death Rate during the COVID-19 Pandemic. *Healthcare* **8**, 339 (2020).

37. van Holm, E. J., Wyczalkowski, C. K. & Dantzler, P. A. Neighborhood conditions and the initial outbreak of COVID-19: the case of Louisiana. *J. Public Health Oxf. Engl.* fdaa147 (2020) doi:10.1093/pubmed/fdaa147.

38. Oguz, B. U., Shinohara, R. T., Yushkevich, P. A. & Oguz, I. Gradient Boosted Trees for Corrective Learning. *Mach. Learn. Med. Imaging MLMI Workshop* **10541**, 203–211 (2017).

39. Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E. & Petrovski, K. R. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models. *Comput. Biol. Med.* **114**, 103456 (2019).

40. Lim, J. T. *et al.* Estimating direct and spill-over impacts of political elections on COVID-19 transmission using synthetic control methods. *PLoS Comput. Biol.* **17**, e1008959 (2021).

41. Haas, E. J. *et al.* Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *Lancet Lond. Engl.* **397**, 1819–1829 (2021).

Tables:

Table 1: *COVID-19 mortality was used as an outcome variable in the models of this study*

Outcome	Description	Collection Location	Time Variant
COVID-19 Deaths per 1000 individuals	Daily COVID-19 deaths were collected and aggregated to weekly values. Since this data was cumulative new deaths per week was calculated in R using lag. For weeks with negative values of new deaths (due to corrective measures in the data), it was assumed that the next positive value of new deaths accounted for multiple weeks, so this value was split so each week between the negative deaths and the positive deaths had the same number of deaths. This data was then further standardized to each county then evaluated as deaths per 1000 individuals.	CDC	Yes

Table 2: *SafeGraph mobility data to four defined location types provided exposure values in the models of this study*

Exposure	Description	Collection Location	Time Variant
Normalized visits by state scaling to K-12 Schools	SafeGraph data on various school locations throughout the county was standardize by scaling the number of visits by the overall amount of people in Georgia. From there all locations with NAIC's codes pertaining to K-12 schools were aggregated to achieve the mean value per week to be used in the analysis.	SafeGraph	Yes

Normalized visits by state scaling grocery stores	SafeGraph data on various grocery locations throughout the county was standardize by scaling the number of visits by the overall amount of people in Georgia. From there all locations with NAIC's codes pertaining to grocery stores and markets were aggregated to achieve the mean value per week to be used in the analysis.	SafeGraph	Yes
Normalized visits by state scaling to transportation terminals	SafeGraph data on various bus and plane terminal locations throughout the county was standardize by scaling the number of visits by the overall amount of people in Georgia. From there all locations with NAIC's codes pertaining to bus stops, bus terminals, and plane terminals were aggregated to achieve the mean value per week. As only 45 counties had any data in relation to this category this was not included in the final analysis.	SafeGraph	Yes
Normalized visits by state scaling to food service locations	SafeGraph data on various restaurants locations throughout the county was standardize by scaling the number of visits by the overall amount of people in Georgia. From there all locations with NAIC's codes pertaining to restaurants or food service locations were aggregated to achieve the mean value per week to be used in the analysis.	SafeGraph	Yes

Table 3: *Three types of movement variables*

Variable Type	Description
Continuous	Numeric mean of normalized visits by state scaling for each category.
Dichotomous	Value determined by looking at the continuous variable and determining if it increased compared to the previous week. This was done to provide a simple way to see if movement in this category was increase as time went on.

Categorical	Using the initial continuous variable, the data was broken into 3 parts, low medium and high levels of movement. This was determined using the cut function in R breaking the continuous variable into three categories, with low movement being below 1/3 of the categories movement and high movement being above 2/3.

Table 4: *Covariates which were identified as potential cofounders for the effect of human movement on COVID-19 mortality were collected and added to the model to provide a better understanding of the relationship between movement and COVID-19 mortality*

Covariates	Description	Collection Location	Time Variant
Per-Capita Personal Income, 2020	Per-capita personal income provides a way to measure the counties socioeconomic status and directly related to poverty levels. Income has a known effect on health outcomes[19].	US Census	No
Percentage of the Population over 65	COVID-19 and general health outcomes have been shown to have a direct correlation with age. COVID-19 has been shown to disproportionately result in deaths individuals over the age of 65[18].	US Census	No
Median age of County, 2019	As another measurement of age, this variable was removed due to the collinearity effect between this and the percentage of the population over 65. Since older individuals had a higher rate of mortality associated with COVID-19, that was determined to be a better measurement for age then this variable[18].	US Census	No

Percentage of Individuals who voted Trump	Political communication and rhetoric has been shown in previous studies to have an impact on the decision of citizens to practice physical distancing. As those choosing not to participate in preventative measures for COVID-19 are at large republican, this was seen as an adequate measurement of the potential percentage of the county who might not be compiling with guidelines[23].	Georgia State Government	No
Total votes in 2020 Presidential Election	As another measurement of political affiliation, this variable was removed due to the collinearity effect[23].	Georgia State Government	No
Urban Code, 2013	Urbanicity is thought to contribute to the overall COVID-19 case numbers and responses[20,21].	US Census	No
Average Household Size	As a way to measure population density, this variable was included since population density has been shown to have a directly effect on the increases of the amount of COVID-19 cases and deaths[22].	US Census	No
Percentage of Black Individuals	Though we are unable to obtain a direct measure of racism in health care and health outcomes, the percentage of the county who identify as Black is a good indication of this level of racism and disparity[42,43].	US Census	No
Population	The population of counties was used to standardize measurements in counties so they are comparable[18].	US Census	No
Week	As we are using a series dataset, the weekly start and end dates for our time period were determined by the aggregated data from SafeGraph[40].	SafeGraph	Yes
New Vaccinations per 1000 individuals	To indicate the decreasing probability of individuals obtaining COVID-19, the total completed new vaccinations per week (per 1000 people) was used .	CDC & Georgia State Government	Yes

Table 5: *Linear Models Evaluation Statistics*

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*

*denotes significant at an alpha level of at least 0.001 for **β coefficients**

Table 6: *GBT Model performance of different movement variable data types*

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

Figures:

Figure 1: *Study Roadmap*

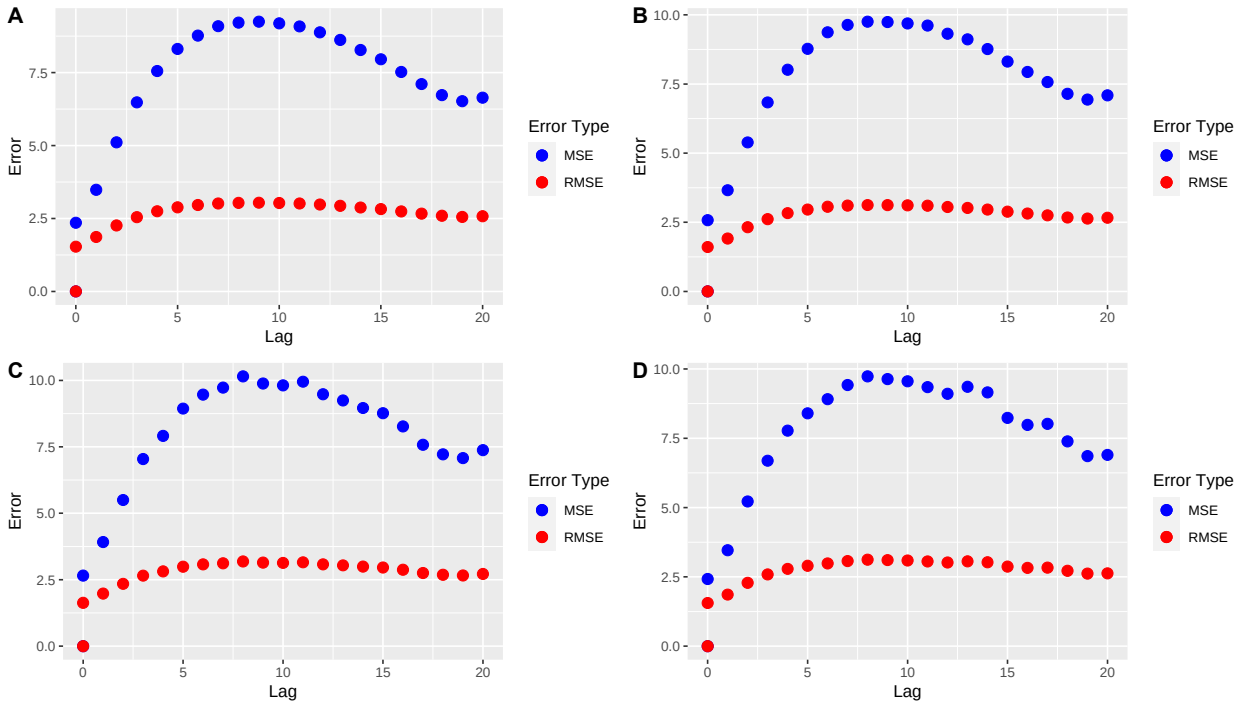

Figure 2: *Assessment of the effect of lagging the movement in GBT models. A- GBT dichotomous movement simple model; B- GBT categorical simple model; C - GBT continuous movement simple model; D: GBT continuous movement full model*

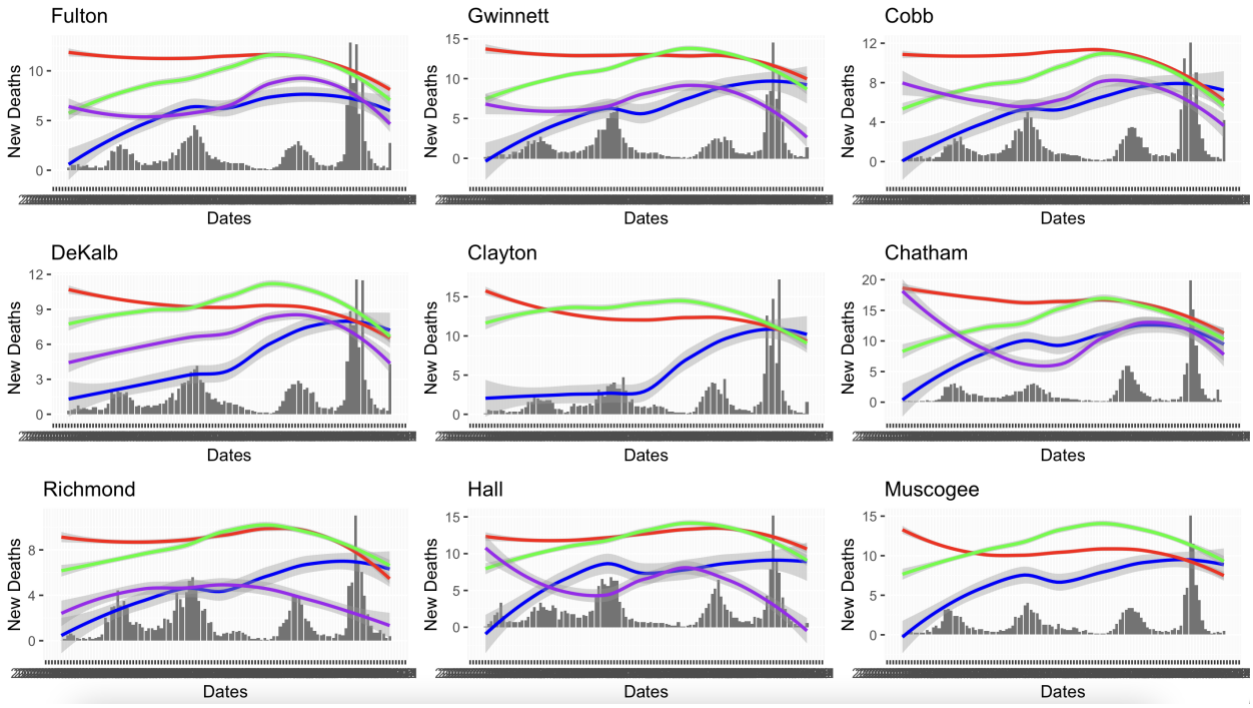

Figure 3: *Histogram of COVID-19 deaths per 1,000 individuals in the top nine Georgia Counties effected by this pandemic. Lines of movement in the county was added to each graph, scaled and smoothed to be seen as the overall effect over time with the following representation of movement catagories: red – education, blue – grocery, green – food services, and purple – transportation*

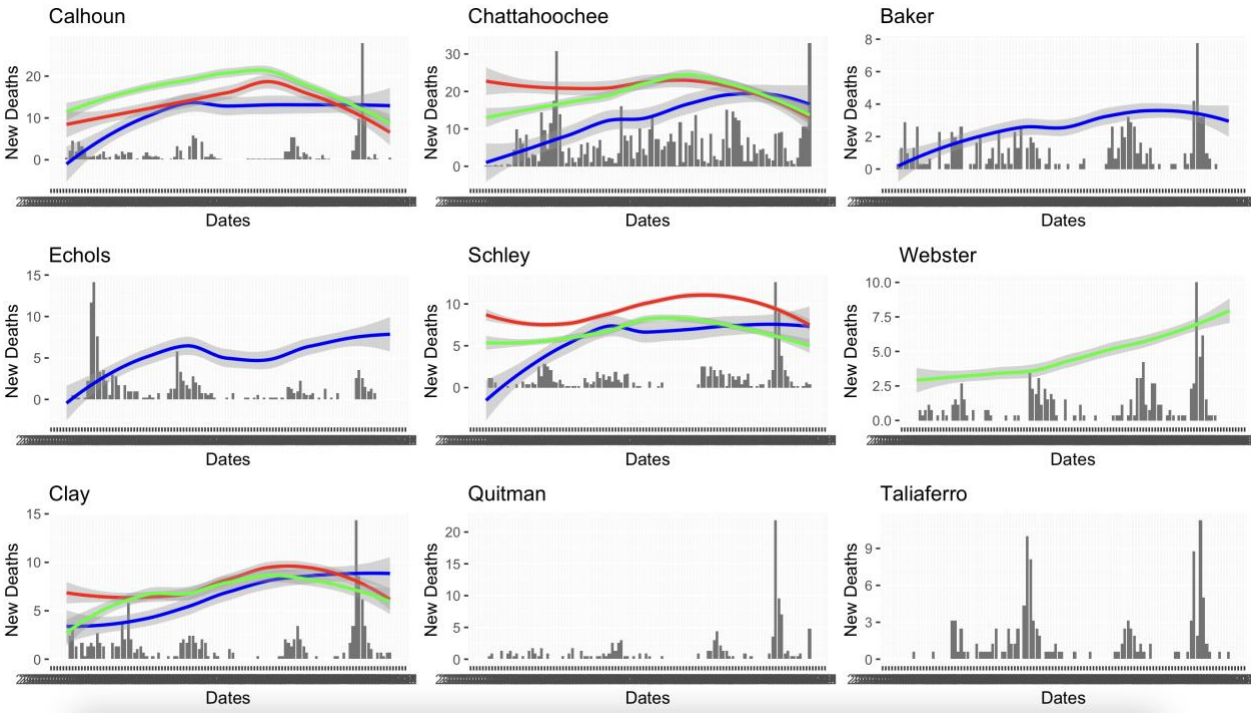

Figure 4: *Histogram of COVID-19 deaths per 1,000 individuals in the bottom nine Georgia Counties effected by this pandemic. Lines of movement in the county was added to each graph, scaled and smoothed to be seen as the overall effect over time with the following representation of movement catagories: red – education, blue – grocery, green – food services, and purple – transportation*

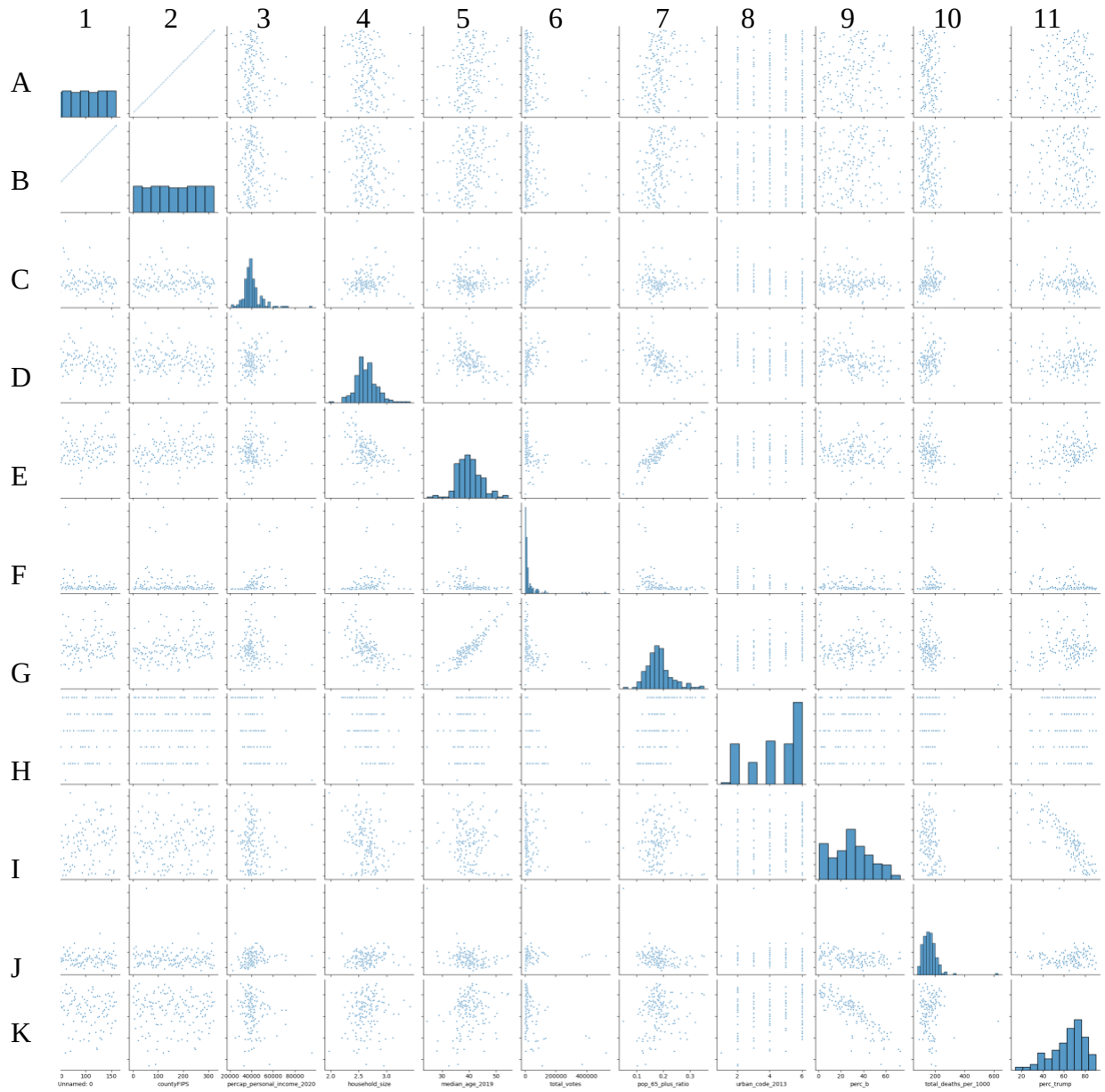

Figure 5: *Comprehensive exploratory graph on all variables to be consider for modeling the effect of human movement on COVID-19 deaths.*

Explanation of Variables in Figure 5

Column # Row Letter	1 A	2 B	3 C	4 D	5 E	6 F	7 G	8 H	9 I	10 J	11 K
Variable	ID	County FIPS	Income	House-hold size	Median Age	Total Votes	% Pop 65+	Urban Code	% Black	Total deaths	% Trump

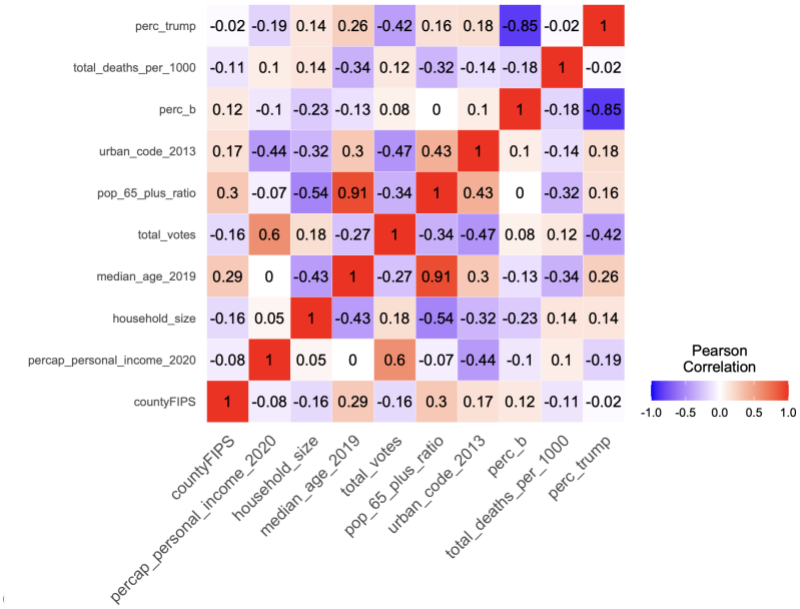

Figure 6: *Variable Pearson correlation heat map used to understand and evaluate potential connections between variables.*

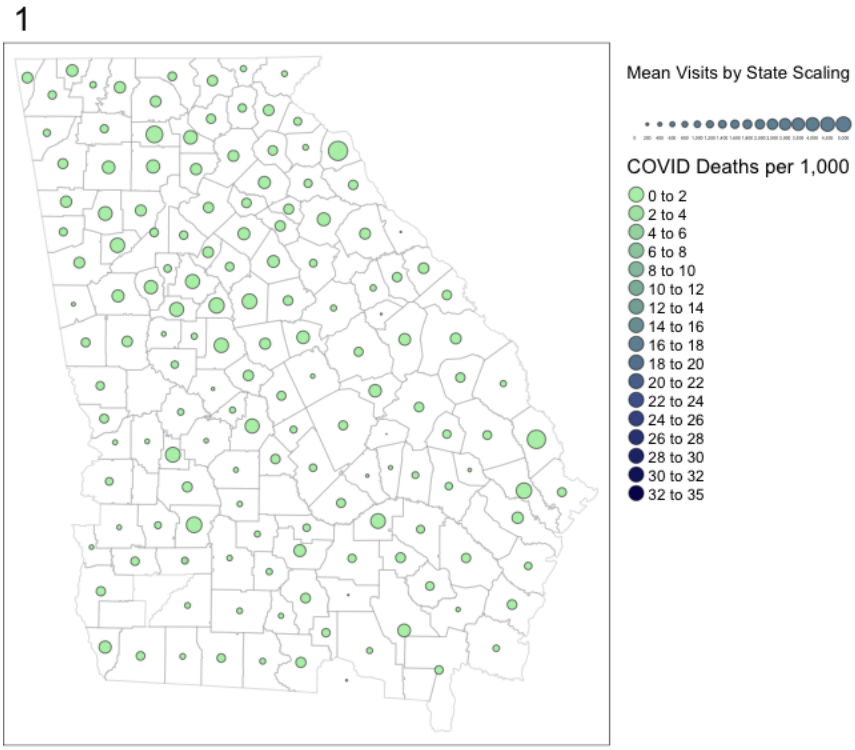

Figure 7: *Frame captured from animated maps showing movement in the grocery category in the first week and the association with COVID-19 mortality*

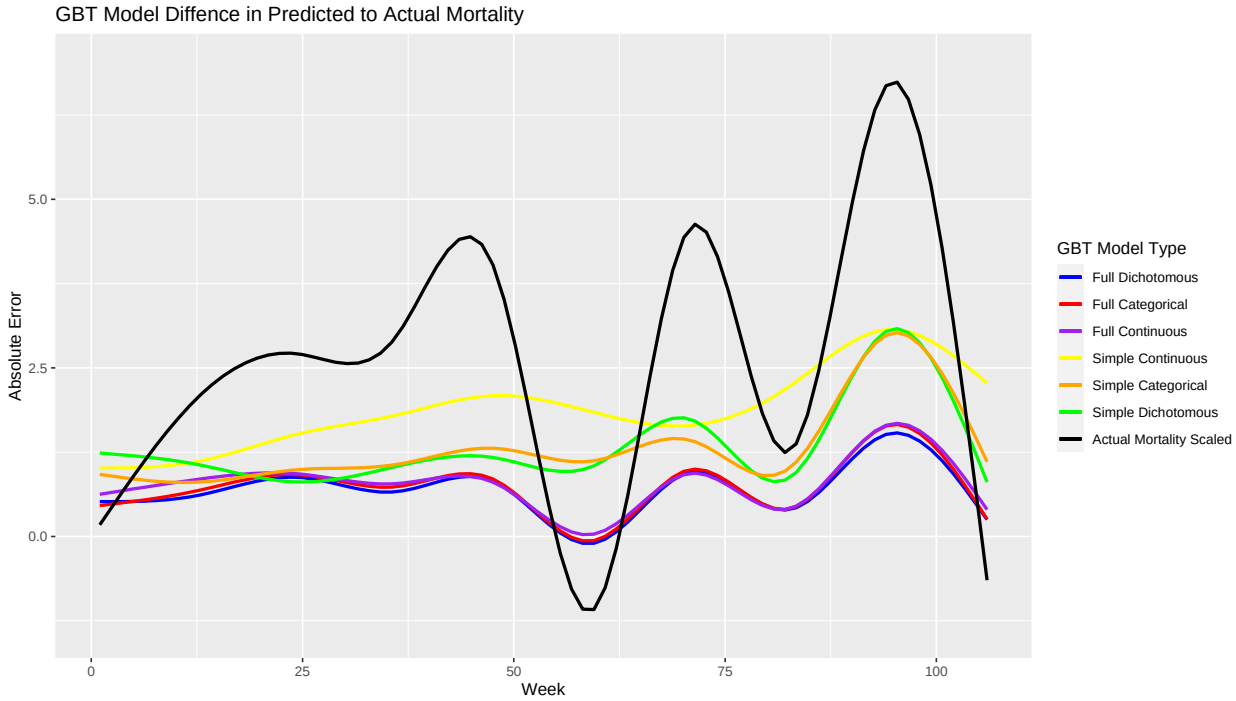

Figure 8: *Absolute difference between predictive and actual for each model in comparison the scaled GA overall mortality per week.*

Appendix 1

Online locations of data collected for analysis in this study:

Covariates	Online Location of data	Main Source	**Time Variant**
COVID-19 Deaths per 1000 individuals	https://apidocs.covidactnow.org/?utm_campaign=API&utm_medium=ppc&utm_source=adwords#historic-data-for-all-states-counties-or-metros	CDC	Yes
Per-Capita Personal Income, 2020	https://apps.bea.gov/itable/?ReqID=70&step=1&acrdn=5	US Census	No
Percentage of the Population over 65	https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html	US Census	No
Median age of County, 2019	https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html	US Census	No
Percentage of Individuals who voted Trump	https://sos.ga.gov/election-data-hub	Georgia State Government	No

Total votes in 2020 Presidential Election	https://sos.ga.gov/election-data-hub	Georgia State Government	No
Urban Code, 2013	https://www.cdc.gov/nchs/data_access/urban_rural.htm	US Census	No
Average Household Size	https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html	US Census	No
Percentage of Black Individuals	https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html	US Census	No
Population	https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html	US Census	No
New Vaccinations per 1000 individuals	https://apidocs.covidactnow.org/?utm_campaign=API&utm_medium=ppc&utm_source=adwords#historic-data-for-all-states-counties-or-metros	CDC & Georgia State Government	Yes

Appendix 2:

https://github.com/erica8494/Thesis_EricaJohnson.git