Predictive Analysis of COVID-19 Cases using Mobility Data

Carson Woods,
CPSC 4180 – Fall 2020

***Abstract*—This work is a preliminary analysis of mobility data provided by Apple and Google to determine if a correlation exists between mobility data requests (from users of apps such as Google and Apple maps) and increased confirmed cases of COVID-19. In this analysis, basic statistical analysis is done to determine correlation as well as visual representation of the data. Additionally, various machine learning models are surveyed by training across data from multiple countries to identify the most useful model type for making predictions on this type of data. In this analysis, the efficacy of various statistical, visual, and machine learning analysis approaches were able to be evaluated. In doing so a correlation was found, however even in the strongest models, it was clear that there is a larger driving force behind spikes in COVID-19 cases than simple mobility data requests.**

***Index Terms*—COVID-19, machine learning, tensorflow, statistical analysis, GRU, LSTM, recurrent neural networks.**

I. Introduction

As a result of the rapid rise of COVID-19 throughout the latter months of 2019 and into 2020, many different techniques and methods have been applied in order to understand the virus and analyze its spread. Various novel visualisations and data sets from sources all over the world have emerged and many attempts to model the disease have been created as a result. One of the most successful and popular visualizations (and corresponding data sets) is the John Hopkins Coronavirus Resource Center which has mapped the global growth of COVID-19 cases by aggregating data from reporting agencies around the world [2]. Additionally, many studies have been done which have attempted to map the disease in a predictable way by trying to find a correlation between the spread of the disease and some other parameter such as hospital bed demand [4]. Other studies have utilized limited mobility data sets to investigate whether how often individuals were traveling somewhere is a driving factor for the spread of COVID-19, however these studies were done with limited data from private data providers [1]. Other studies that focused on mobility took into account mobility restrictions, but were limited in their scope to a single country [6]. While this work also focuses on determining if there is a correlation between mobility data and increases in COVID-19 cases, the data used in the following analysis is publicly available and is provided by Apple and Google maps and covers a much longer stretch of time than other studies [3, 5].

II. Materials

The approach taken in this work is a straightforward one. Apple and Google's Mobility data sets were used to determine if there was a positive correlation between the amount of directional data requested and the number of COVID-19 cases. For the sake of this analysis "directional data" and "mobility data" are the same and refer to an individual or (in the case of the whole data sets) a group of individuals requesting directions using one of the map providers (Apple Maps or Google Maps).

A. COVID-19 Case Data

The Johns Hopkins data set was a single CSV file, that contained time series data of confirmed cases for 58 countries around the world. The case data was structured cumulatively so that each day included all cases from previous days [2]. This data acted as the source for determining increases or changes in COVID-19 cases for this analysis.

B. Apple Mobility Data

Apple and Google took different approaches when collecting their data, but the general thought process was the same. Both Google and Apple took mobility data from a standard day (prior to lock downs and widespread COVID-19 cases) and set that mobility data as the baseline that all other days would be measured against [3, 5]. Days following the baseline were given a number that corresponds to the percent change (increase/decrease) of mobility data requested. How this data was represented was the first (and smallest) difference between the data sets. Apple's approach was to set the baseline at 100 and future days would either be more or less than that; for example, if one day had 5% more requests, than the baseline the number recorded would be 105 [3]. If it were 5% less, then the number would be 95. Apple's data set also took a fairly straightforward to data representation. It broke up the data into tiers, the available geographic tiers were:

- **Countries** - Ex: United States
- **Sub-Regions** - Ex: Tennessee
- **Counties** - Ex: Hamilton County
- **Cities** - Ex: Chattanooga

and each tier had all the data for the tiers that fell within it [3]. For example, if the United States was listed as the country, that data would include all mobility data for all sub-regions that fell within the US (all of the states), and so forth. This allows for analysis at various different levels of locality. Finally, the data itself was divided up into subsections depending on the type of directions being requested. Apple provided 3 data categories: driving directions, walking directions, and public transit directions. Driving and walking were provided for every region, however many regions did not have transit data as public transit is not widely available.

C. Google Mobility Data

Google's approach to structuring their data set was similar, but they started at a baseline of 0, and went positive or negative to represent changes in the quantity of mobility data requested. A 5% increase or decrease in the Google mobility data would result in 5 or -5 being recorded respectively [5]. Google's geographical division of their data was similar to Apple's; however the naming scheme was different. Google's Data used the following regional breakdown:

- **Country or Region** - Ex: United States
- **Sub-Region 1** - Ex: Tennessee
- **Sub Region 2** - Ex: Hamilton County
- **Metro Area** - Ex: Chattanooga

In the end, Google's data was geographically divided almost identically to Apple's data set, but used different labels. The real differences from Apple's data came in the data itself. Unlike the transit type division that Apple used, Google used destination type as the category for dividing the data [5]. Google's data included requests for directions to the following categories of destinations:

- Retail and Recreation
- Grocery and Pharmacy
- Parks
- Transit Stations
- Workplaces
- Residential

rather than dividing the data up by transportation type [5].

III. Methods

While both mobility data sets express similar types of data, their formatting, structure, and overall consistency is quite different. Before analysis could be done, a great deal of pre-processing was required to unify the structure of data to a single standardized format. Once this was complete analysis could begin. In the following section, the pre-processing steps will be explained, as well as an in-detail description of the analysis that was undertaken.

A. Pre-Processing

As mentioned, the format of these data sets was not the same. The first step was deciding on a format for the final data. Upon examining the data set, it was decided to break the data up into countries and to do analysis on a per-country basis. A more detailed explanation of this decision is included in the discussion section. The pre-processing was done by iterating through each data set and separating out the data on a country per country basis. The indexes and labels for each data set are modified to match each other; specifically the date formats for each data set were slightly different, so they were modified to match. Once the data was separated, data from each country was combined into a unified, miniature data set for each country containing data from all three data sets (Apple mobility, Google mobility, and John Hopkins COVID-19 data). While name based filtering during the analysis step was a consideration, it ended up being much simpler to do the pre-processing in advance, then run analysis on each country individually. Additionally, by doing all the processing in advance, no pre-processing steps were repeated unnecessarily which cut down on run-time dramatically.

B. Analysis

Once the data was pre-processed, per-country analysis was done on the newly separated data. The analysis was done using two different methods. First the data was visualized. Each country had a series of graphs generated for it, where each graph was a visualization of a different data set and its impact on COVID-19 cases over time. This was done with time in days since the start of the data set being placed on the x-axis of the graph and COVID-19 cases being placed on the y-axis. The graph was represented as a scatter plot, with the size of each point on the scatter plot scaling in size to represent a week's mobility data average. The larger points on the graph, the more mobility data was requested in that country during that week, the smaller the point, the less data was requested. This allowed for representation of a data set with more than 2 dimensions to represent. Once this was done and the graphs were saved, there was an attempt to determine if there was a reasonable correlation between mobility data requests and COVID-19 cases using statistical methods. To determine correlation, a linear regression model was created from each country's data.

The majority of the analysis done for this project involved using machine learning to identify a trend (and test forecasts) on the data sets. Because COVID-19 cases are linked to those cases that came before it, a recurrent neural network was selected as the model that would most likely fit this type of data. Recurrent neural networks (or RNN's) are particularly good at analysing sequential data, and it was a natural fit for this time-series data. The machine learning framework used by this analysis is Tensorflow, which provides 3 built-in RNN layers for building neural networks: a simple RNN layer, a long short-term memory (LSTM) layer, and a gated recurrent unit (GRU) layer. All three types of RNN layers were tested using the same process. The models were all created so that they contained only two layers, the RNN layer of choice (SimpleRNN, LSTM, or GRU) and a Tensorflow Dense layer that functioned as the output layer for the model. This purpose of this analysis and testing wasn't to create the ”best” model for making these predictions, but instead it was designed to determine if machine learning was a useful technique in trying to make predictions at all on these particular data sets. After constructing the models, the data sets were split up into distinctive training, validation, and testing subsets (used to later validate the results), and the model was fit to the training data over 500 training epochs. For all models, the Adam optimizer was used, and a mean-squared loss function were chosen from Tensorflow's provided optimizer and loss functions.

IV. Results

As expected, the results for this analysis varied dramatically depending on the country whose data was being analyzed, however there were particularly good and bad examples of where these analysis models were more and less effective respectively. In the following section, some of these more prominent results are outlined, as well as averages for all model types across all countries included in the analysis.

A. Visualization

While the visualizations were not able to definitively determine correlation, it did offer insight into whether continuing analysis with this data set was worthwhile. Of the generated visualizations, some showed a stronger visually obvious correlation than others, but the most striking visualizations were the very good ones, or the very bad ones. As seen in figure 1, countries like Albania, had a very good correlation that clearly shows cases increasing more rapidly when driving mobility requests increase. By comparison, the worst visualizations came from countries with a worse or less obvious correlation. The United Kingdom is one surprising example of this, as can be seen in figure 2. The image seems to show the exact opposite correlation that Albania exhibits, where the increase in driving directions requested actually corresponded to a decrease in the rate of new cases.

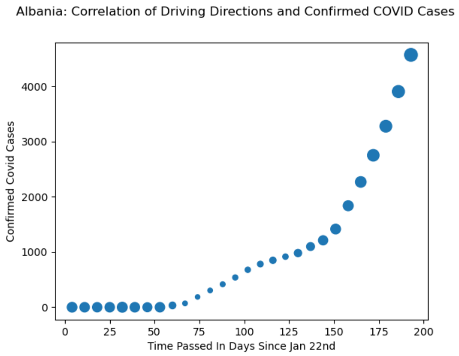

Fig. 1. Cases increase slower in Albania when driving directions are requested less.

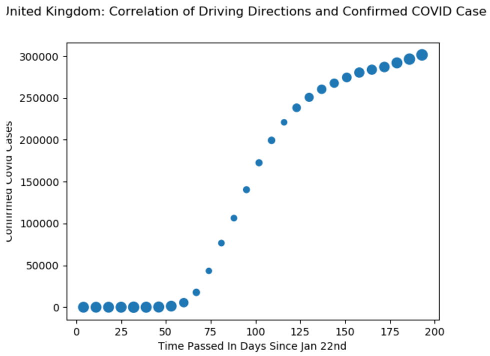

Fig. 2. The United Kingdom's data shows the highest rate of cases when mobility requests are lowest.

B. Linear Regression Results

After performing the linear regression, it was found that for almost every single country, linear regression is a less useful model for making predictions. While a more detailed explanation is included in the discussion section as to why likely is, it's worth noting now that this wasn't an altogether unexpected result. When measuring the R-squared value for goodness of fit (the closer to 1 the R-squared value is the better the fit of the model), most countries had values that were too low to make meaningful predictions. While it is not required for the R-squared value to be almost exactly 1 for the model to be useful, only 5 countries even had a single r-squared value over .5, with Sweden's driving mobility data (sourced from Apple) having the highest at approximately .737. In most cases the R-squared value for the regressions (regardless of the type/source of data) were below .2, with Vietnam's driving data being the lowest with approximately 0. After analyzing the results, the highest average R-squared value for any one data type came from Google's residential data set, resulting in an average value of 0.189. The lowest regression average was Google's workplace data set at .0899. Despite all of this, the linear regression model is not without room for improvement. Many of the models that were generated were statistically significant (p-value ≤ 0.05) and even Vietnam's models, which was the country with the worst R-squared value, had p-values of $p \leq 0.001$.

C. Machine Learning Results

The machine learning results were not nearly as straightforward, however they also showed that machine learning is generally a much more effective model for making predictions. Generally, it was found that loss for all three models tested was quite reasonable on the training data, with the SimpleRNN generally performing the worst of the three (it had the highest loss) and the GRU model had the best performance on average (lowest average loss). This trend held true when evaluating the models on the validation and testing subsets of the data. On average, the SimpleRNN model had an average loss of 0.7812 and 6.0155 on validation data and testing data respectively. By comparison, the LSTM model had an average loss of 0.4589 and 4.9787 for validation data and testing data respectively. And finally the GRU model slightly outperformed the LSTM model on unseen data with loss values of 0.4576 and 4.8333 for validation data and testing data respectively.

V. Discussion

This sections covers a discussion of the analysis results, as well as inferences as to why particular results are more effective than others at forecasting future COVID-19 cases. The machine learning results are the most effective for trying to make predictions, however its also worth looking at why the linear regression performed so poorly and why a poor result was expected from that method of analysis.

The linear regression is a relatively simple model which attempts to find the best linear function that maps the results to the data. As a result, this was expected to be a bad model when compared to other, more complex attempts to map a correlation between mobility data and COVID-19 cases. The primary reason for this failure is because it simply isn't a linear correlation. The COVID-19 case data is clearly not linear and even if the mobility data was also not linear, the linear regression would not do a good job modeling this behavior (linear inputs to non-linear outputs). Despite these shortcomings, many of the linear regression models were statistically significant, despite not being particularly useful for predictions. This indicates that, while there is a link between COVID-19 cases and mobility data, it is more likely that there is another factor involved that is driving changes in COVID-19 cases. It is also likely that whatever other factor or factors is influencing COVID-19 cases, exhibits a high level of multicollinearity with mobility data. While it is possible that a better model could have been created through higher-order statistical regression or targeting different independent variables, it likely would have still been less effective than a machine learning alternative, since a regression model simply can't take into effect as many variables or contributing factors as a machine learning model can. One of the most significant differences between the machine learning model and the regression model is that the machine learning model uses every part of the data sets made available to it simultaneously, where the regression model only used one type of mobility data at a time. Even if the regression model were a better fit, the machine learning model can train on a more comprehensive selection of data, as well as it can be scaled to other data if it becomes available.

Of all the machine learning models tested in this analysis, the LSTM and GRU machine learning models were the most effective, and performed nearly identically, with the GRU model being slightly better on average. Unfortunately, despite the models showing excellent loss while training, the testing loss being much higher likely implies some data over-fitting in the model. While this over-fitting isn't ideal, it doesn't mean that the model isn't a viable option for making forecasts of this nature; there are many possible improvements that could be made.

One improvement that could improve the model is simply expanding the data set. As it stands now, only around 250 days of data were included in the mobility data; because of the smaller data set and relatively uniform mobility data, over-fitting is likely made worse through over training the model. One possible way to combat this would be to increase the size of the data set. This will naturally happen as COVID-19 continues to be a public health issue, however it takes significant time for the data set to grow enough to overcome this issue. Another option for improving this model is to include different types of data in the analysis. While it is clear that there is a correlation between mobility data and COVID-19 cases, it is also clear that mobility data is not the driving or limiting factor for the spread of the disease. These types of predictions would likely be more effective if, for example, lockdown and mask regulations was also included into the data set in some capacity. This would likely be more effective since the mobility data would complement the much more authoritative government issued restrictions. Mobility data alone doesn't indicate if people are staying safe while traveling or at their destination, it only shows if people are requesting directions to a location. Despite this it could create a very effective model by providing insights into if people are following restrictions in their area and seeing how those trends collectively map to increases or decreases in COVID-19 cases. Another possible change that could be made to this model in future work is changing the COVID-19 data to not be cumulative. Since mobility data is likely able to offer insights into the rate of change of COVID-19 cases, it might be more effective to train on the number of new cases each day, rather than cumulative cases since cumulative cases could add weight to countries with many cases or larger populations, even if they have stopped the spread later. By using new cases rather than cumulative, the model could be trained on the rate of change rather than the actual amount of cases total. The final, and most obvious, improvement that could be made is model refinement. The models used in this analysis were extremely basic, and could easily be improved upon. Adding more layers, varying loss functions, mixing RNN types, etc. are all different ways that the model could potentially be changed to improve results. While this analysis did prove the viability of making predictions with models such as these, it is clear that many improvements could be made to increase performance without needing to change the data set being trained on in any way.

VI. Conclusion

In conclusion, the results of this analysis clearly show a correlation between mobility data and COVID-19 cases. This analysis shows that there can be reasonable predictions made based on the mobility data provided by Apple and Google, however it has also identified many areas of improvement that future work and additional analysis could make. Future work should focus on improving the machine learning models and expanding the data sets used. By increasing the size of the mobility data sets, including non-mobility data, and improving the machine learning model architectures, the quality of the predictions made by models like those included in this analysis can be dramatically improved. While it would be ideal to be able to properly predict where high numbers of new COVID-19 cases are going to appear based solely on mobility data, the reality is that a holistic approach that focuses on many factors of disease spreading would be much more likely to produce favorable results.

References

[1] Hamada S Badr et al. "Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study". In: *The Lancet Infectious Diseases* 20 (2020), pp. 1247–1254. DOI: https://doi.org/10.1016/S1473-3099(20)30553-3. URL: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30553-3/fulltext.

[2] Ensheng Dong, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet Infectious Diseases* 20.5 (2020), pp. 533–534. ISSN: 1473-3099. DOI: https://doi.org/10.1016/S1473-3099(20)30120-1. URL: http://www.sciencedirect.com/science/article/pii/S1473309920301201.

[3] Apple Inc. *Mobility Trends Reports*. 2020. URL: https://covid19.apple.com/mobility.

[4] B. Ivorra et al. “Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China”. In: *Communications in Nonlinear Science and Numerical Simulation* 88 (2020), p. 105303. ISSN: 1007-5704. DOI: https://doi.org/10.1016/j.cnsns.2020.105303. URL: http://www.sciencedirect.com/science/article/pii/S1007570420301350.

[5] Google LLC. *Google COVID-19 Community Mobility Reports*. URL: https://www.google.com/covid19/mobility/.

[6] Ying Zhou et al. “Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data”. In: *The Lancet Digital Health* 2.8 (2020), e417–e424. ISSN: 2589-7500. DOI: https://doi.org/10.1016/S2589-7500(20)30165-5. URL: http://www.sciencedirect.com/science/article/pii/S2589750020301655.