Bridging the Gap: Enhancing COVID-19 Epidemic Forecasting by Integrating Social Mobility Dynamics into Time Series Models

Mehul Rastogi
mehulrastogi@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Akshat Karwa
akarwa7@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

1 INTRODUCTION

In this project, we aim to enhance the precision of COVID-19 epidemic forecasting by integrating social mobility dynamics into foundational time series models. By acknowledging the significant influence of human movement and social distancing on viral transmission, we will conduct an analysis of historical COVID-19 data in conjunction with social mobility data. This will allow us to investigate the correlations between mobility patterns and infection rates.

Our methodology includes the implementation of diverse time series forecasting models to simulate the spread of COVID-19. We will evaluate the models' effectiveness in capturing the role of social mobility in transmission dynamics. Moreover, we will integrate comprehensive visualisations to elucidate the effects of varying social distancing measures across different U.S. states. Subsequently, our objective also involves critically assessing the advantages of foundational time series epidemiological models in replicating the observed transmission patterns of COVID-19. Leveraging real-world social mobility data and COVID-19 time series data for the United States, we will examine how well existing models account for the impacts of human movement and social distancing.

Overall, we plan to perform a rigorous analysis of models such as ARIMA, LSTM, Facebook Prophet, SIR Models, and Hidden Markov Models, contrasting their forecasts with the actual trajectory of COVID-19 spread thereby identifying both their advantages and limitations. By incorporating social mobility data, we aspire to refine these models and enhance their capacity to accurately simulate viral transmission dynamics across various US states.

2 PROBLEM DEFINITION

The COVID-19 pandemic has underscored the critical importance of effective epidemic forecasting. Accurate epidemic forecasting can help inform public health policies and resource allocation. Traditional forecasting models have primarily focused on epidemiological parameters, such as infection rates and recovery times, often overlooking the role of social mobility dynamics in influencing viral transmission. Human movement patterns are shaped by a variety of factors that include social behaviour, government policies, economic activities, and cultural norms. These dynamics can influence how individuals interact, travel, and gather, ultimately affecting the transmission of infectious diseases like COVID-19. As social distancing measures vary across regions and time, there is a compelling need to understand how these factors interact with epidemiological data to enhance the accuracy of forecasts. In this context, the problem we aim to address is twofold.

First, we seek to identify and quantify the relationship between social mobility and COVID-19 infection rates using historical data. By correlating mobility data and COVID spread data, we aim to uncover how fluctuations in human movement influenced the trajectory of COVID-19 cases.

Second, we plan to evaluate how the effectiveness of time series epidemiological models changes after incorporating social mobility data when simulating the spread of diseases. This involves contrasting the predictions of existing models with the actual infection trajectories observed during the pandemic with and without social mobility data integration.

Overall, we aim to identify the strengths and weaknesses of these models and gain insights into their predictive capabilities. Our goal is to refine these models to enhance their utility in real-world epidemic forecasting so that they can aid public health officials in making informed decisions to mitigate the spread of infectious diseases.

3 LITERATURE SURVEY

3.1 Research Paper 1

Exploring the relationship between mobility and COVID-19 infection rates for the second peak in the United States using phase-wise association [1]

The paper has accessed and understood infection and mobility data across different periods like lock downs and re-openings. This paper has also been able to understand and study the different phases of the spread thoroughly. Also, mobility data has been collected from sources like Apple, Facebook, and Google. To make this even more powerful, they've also understood how changes in public perception and government policies have impacted the spread of the virus. The paper also goes over a multitude of US geographies ensuring a cohesive study. However, some drawbacks are that the Google, Facebook, and Apple data isn't representative of actual mobility. If we'd try to use this approach for less tech-y countries, it would

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

lead to a failure. Also, the paper doesn't take into consideration the fact that "correlation doesn't imply causation" bias. Along with mobility, other facts like mask mandates and city-wide precautionary measures should also be measured. There are some interesting approaches used by the paper. However, we want to expand even further and use different time series models along with mobility data to understand how we can map and predict infection spread effectively.

3.2 Research Paper 2

The Twitter Social Mobility Index: Measuring Social Distancing Practices from Geolocated Tweets [2]

This paper presents the Twitter Social Mobility Index which is a measure of social distancing and travel patterns and is derived from publicly geolocated Twitter data. This index uses publicly available geolocated Twitter data which we will be utilising as a data source for our project. The mobility movement is calculated by computing how much a person has travelled in a given week based on the standard deviation of their tweet locations which are geolocated. This will help us in capturing both the area and regularity of travel. Moreover, the index provides mobility measures for the entire US, individual states, and major cities. The temporal coverage in the data is from January 1, 2019 to April 27, 2020 which allows for the comparison of mobility patterns before and during the COVID-19 pandemic. The paper quantifies mobility reductions after social distancing measures were implemented, both at group and individual levels. The authors also analyse correlations between mobility, COVID-19 case rates, and other relevant factors like state size, homelessness, unemployment, etc. Through real-time updates, the authors maintain a website to provide ongoing updates to the mobility index during the pandemic. For our proposed project, this Twitter Social Mobility Index would serve as a valuable data source for social mobility patterns at various geographic levels. It would also be a benchmark to compare against other mobility data sources, and a way to incorporate real-time, publicly available mobility data into COVID-19 forecasting models. The authors correlate mobility patterns with tweet content and this paper would be a significant resource offering us unique advantages for integrating social dynamics into epidemic forecasting models.

3.3 Research Paper 3

Mobility network models of COVID-19 explain inequities and inform reopening [3]

This paper focuses on individual data at specific inflection points like restaurants, bars, etc. The data used is anonymous. It's basically mobile anonymous data collected via SafeGraph. The paper dissects mobility patterns based on difference in socio-economic groups. They are able to come to a conclusion that infection spread is more volatile in lower-income areas. However, this paper doesn't use time series models and focuses on the network based approach of humans. The approach is quite interesting. However, we believe that using mobility data along with time-series models can lead to much more rewarding results than just difference in covid-spread based on income-inequalities.

3.4 Research Paper 4

Epidemic forecasting based on mobility patterns: an approach and experimental evaluation on COVID-19 data [4]

This main premise that we are interested in this paper is that LSTM models along with real-time mobility data is much accurate than other methodologies in forecasting the spread of an infection. Mobility is measured in terms of spatial and temporal patterns. The only issue with such mobility requirements is that the data required is extremely complex. It's very hard to firstly, collect this data and secondly, the processing required to iteratively construct these spatial and temporal patterns is extremely hard and ineffective. But the approach of using LSTM models with mobility data is very novel, and we will bridge idea from the paper in our implementation.

3.5 Research Paper 5

Trending on the use of Google mobility data in COVID-19 mathematical models [5]

This paper highlights how time-series models like SIR can be used with mobility data to get more accurate infection spread predictions. Real-world data can be mapped in a much more efficient manner. In the paper, the Google mobility data is used with the mathematical models to predict infection spread. There's no major alarming drawback in this paper. However, it can be of great help for our implementation since we can draw some parallels between this paper and our implementation. Essentially, we will be drawing mobility data from many sources to maintain accuracy and also calibrate each of the mathematical models with mobility data efficiently to ensure accurate spread data can be predicted.

3.6 Research Paper 6

COVID-19 Pandemic Prediction using Time Series Forecasting Models [6]

This paper evaluates time series forecasting models for predicting COVID-19 cases, utilizing data from January 22, 2020 to May 20, 2020 for the United States. It evaluates two time-series forecasting models: ARIMA (Autoregressive Integrated Moving Average) and Facebook Prophet. These models are used to forecast confirmed, active, recovered, and death cases. The performance is evaluated using diverse metrics like MAE, RMSE, RRSE, and MAPE. We will build on what is done in this paper and implement time series models. We will also utilise these evaluation metrics and similarly compare model performance. We also plan on extending our research to integrate with factors like lock-downs, social distancing, population density, and healthcare capacity. We will utilize this paper's methodology and insights to implement similar models, which will potentially assist us in understanding disease trends and informing policy decisions.

Weaknesses include considering limited variables, assuming stationarity, having sensitivity to initial conditions, a short-term focus, and the lack of epidemiological context.

3.7 Research Paper 7

Exploring the effects of non-pharmaceutical interventions on COVID-19 transmission through machine learning and modeling approaches. [7]

This paper focuses on interweaving Machine Learning techniques with Non-pharmaceutical interventions (NPI) like social distancing, mask wearing, and travel restrictions. Different approaches like SIR and ML-based models are used along with this NPI data to accurately predict the spread of the virus. NPI is incorporated into time series models like LSTM, ARIMA etc. NPI information is also used as independent variables in regression models. So, this paper uses such information to increase the accuracy of both mathematical as well as ML models. However, since collecting NPI data at a granular level is extremely hard, improving the accuracy is very hard. We will draw parallels from this research in-our implementation and will improve the quality of our NPI data.

3.8 Research Paper 8

Spatial-Temporal Relationship Between Population Mobility and COVID-19 Outbreaks in South Carolina: Time Series Forecasting Analysis. [8]

This paper looks at the relationship between population mobility and COVID-19 outbreaks in South Carolina. It utilizes a Poisson count time series model and finds that population mobility was positively associated with COVID-19. The paper also claims that state-level models had a much higher prediction accuracy than county-level models. We will utilize this paper's approach of incorporating the Twitter-based mobility data and examining different region predictions. Limitation of this paper is only using Twitter data and focusing on one state. We will consider different regions along with mobility factors and utilize a combination of datasets.

3.9 Similarity and Different between Papers

The primary similarity among most papers lies in their utilization of mobility data which is aggregated from several sources. These papers aim to juxtapose this mobility data with time-series or machine learning models. Some studies also incorporate additional nuances of mobility, for example mass mask mandates, etc., in order to further enhance predictive accuracy. While some papers focus on national trends across the United States, others delve into more localized geographies like counties or states. This diverse range of approaches makes us think deeper about whether we should generalize findings for the entire country or concentrate on specific niche areas. Different types of models such as linear regression, SIR, and LSTM—are employed alongside the mobility data across these studies. Ultimately, despite their diverse methodologies, we see that all the papers strive to improve the predictive accuracy of models discussed in them by integrating mobility and non-pharmaceutical intervention (NPI) data aiming for more precise predictions.

4 ALGORITHMS, TECHNIQUES, AND MODELS

4.1 ARIMA (Auto-Regressive Integrated Moving Average)

This time-series model uses the following three components for predictions:

Auto-Regressive (AR)   Integrated (I)   Moving Average (MV)

4.2 LSTM (Long Short-Term Memory Networks)

A LSTM would be utilized to take a Recurrent Neural Network (RNN) based approach to simulate the spread of COVID.

4.3 Facebook Prophet

This is a Open Source model by Meta to forecast disease spread efficiently.

4.4 SIR Models

We will use SIR (Susceptible, Infected, Recovered) models and consider using more advanced alternatives like SEIR, SIRS, SEIRS, and SIRV models. We will most probably utilize a SIRV (Susceptible, Infected, Recovered, & Vaccinated) model for our project.

4.5 Hidden Markov Models (HMM)

A HMM would be utilized to predict the spread of a virus based on historical epidemic data.

4.6 Global Epidemic and Mobility Model (GLEaM) [10]

This is another model we would utilize to work with human mobility data to accurately model virus spread.

5 DATA SOURCES

Some of the datasets that we plan on utilizing our mentioned below. As we move forward with our project, we will look further into more time series data sources that could complement our project's needs.

5.1 COVID-19 Twitter Social Mobility Data [9]

The dataset contains the current index and longitudinal data for several cities, and all the states in the United States.

The longitudinal data has shape $(209, 161)$ which includes mobility data for more than 200 days for 100 most populated cities, all states, 5 regions (Northeast, Midwest, South, Central, West), and 2 territories (Puerto Rico and Virgin Islands). The current index data has shape $(154, 8)$ which includes mobility data for different locations and across different features such as mobility before and after distancing, reduction in mobility, number of users and number of unique tweets.

5.2 Google Community Mobility Reports [11]

This dataset tracks changes in movement patterns and provides insights on how people visit locations like retail, parks, and workplaces varied by geography.

5.3 COVID-19 Data by Our World in Data [12]

This dataset has been collected from the WHO COVID dashboard and it contains data of shape $(1674, 68)$ for the United States,. There is data for 1674 days and there are 68 different features.

5.4 COVID-19 Data Repository by the Center for Systems Science and Engineering [13]

This dataset contains city-wise time series data for the spread of COVID from January 22, 2020 to March 9, 2023 along with longitude and latitude data.

6 EVALUATION AND TESTING

There are some model performance metrics that we are particularly interested in:

(1) Mean Absolute Error (MAE)
(2) Root Mean Squared Error (RMSE)
(3) Mean Absolute Percentage Error (MAPE)
(4) Relative Root Squared Error (RRSE)

Next, we're interested in comparing the model performance with existing time series models. We want to compare our model performance that has mobility data accounted for, and see how does it compare to the naive implementation of SIR, LSTM, or other models.
Furthermore, we want to implement k-fold-cross-validation to ensure that there is no over-fitting in our models. Some of the visualization that we will create are:

(1) Predicted vs. Actual Case Counts
(2) Impact of mobility on course of COVID spread
(3) Trajectory of spread across different locations

We will further implement real-world testing in which we would compare the model projections with occurrences witnessed in the real-world witnessed.

7 PROJECT GOALS AND IMPACT

7.1 Goals

In this project, we aim to enhance the accuracy of COVID-19 epidemic forecasting by integrating social mobility dynamics into traditional time series models. We plan to achieve the following:

Quantify the Relationship between Mobility and COVID-19 Spread: By analyzing COVID-19 infection rates alongside mobility data (e.g., from Google, Twitter), our objective is to uncover correlations between social behavior patterns and infection trends. By doing this, we will be able to identify key mobility indicators that influence virus transmission.

Evaluating and Enhancing Forecasting Models: We will implement and compare diverse time series models (e.g., ARIMA, LSTM, SIR, Facebook Prophet, Hidden Markov Models) to understand how accurately they can track and predict COVID-19 spread with and without the integration of the mobility parameter. Moreover, we will refine these models to enhance their accuracy and predictive capabilities.

Visualizing the Impact of Mobility on Epidemic Dynamics: We will develop comprehensive visualizations to illustrate how variations in mobility patterns influence infection rates across different U.S. states, major cities, and different regions.

Assessing Model Performance and Utility: We will critically analyze the advantages and limitations of each model in replicating real-world scenarios. Furthermore, we will put forward the similarities and differences between each model and suggest improvements to enhance their utility for future epidemic forecasting.

7.2 Impact

With the integration of a social mobility dynamic, our project will not only improve the accuracy of current state of the art models, but will also provide policymakers with actionable insights. With the influence of human movement patterns on infection spread quantified, public health officials can make much more informed decisions. These decisions will be based on numerical evidence regarding how to implement or relax social distancing measures. This will help mitigate the impact of future outbreaks. Our project will also bridge the gap between traditional epidemiological models and real-world social dynamics.

7.3 Stretch Goals

In addition to integrating social mobility, we aim to narrow the gap further by incorporating features such as smoking history, diabetes prevalence, climate factors, economic status, population density, and more, to enhance our models further. We aspire to contribute to the development of more reliable and responsive epidemic forecasting tools.

8 TIMELINE

All the work will be done by both the individuals - Mehul Rastogi and Akshat Karwa

(1) **Phase 1 (10/09 - 10/19)**: Data Cleaning and Preprocessing

(2) **Phase 2 (10/19 - 10/24)**: Exploratory Data Analysis (EDA), Visualizations, and Key Feature Identification

(3) **Phase 3 (10/24 - 11/14)**: Model Implementation with Baseline Parameters

(4) **Phase 4 (11/14 - 11/19)**: Hyperparameter Tuning if Needed and Integrate all the models with Mobility Data

(5) **Phase 5 (11/19 - 11/29)**: Run the Model in Parallel to Evaluate Performance Across Different Metrics

(6) **Phase 6 (11/29 - 12/03)**: Final Report Compilation and Documentation

REFERENCES

[1] Gottumukkala, R., et al. (2021). Exploring the relationship between mobility and COVID-19 infection rates for the second peak in the United States using phase-wise association. *BMC Public Health, 21*(1). https://doi.org/10.1186/s12889-021-11657-0

[2] Xu, P., Dredze, M., & Broniatowski, D. A. (2020). The Twitter Social Mobility Index: Measuring Social Distancing Practices With Geolocated Tweets. *Journal of Medical Internet Research, 22*(12), e21499. https://doi.org/10.2196/21499

[3] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. Nature 589, 7840 (Jan. 2021), 82–87. DOI:https://doi.org/10.1038/s41586-020-2923-3

[4] Pizzuti, C., Rossetti, G., & Spena, M. R. (2022). Epidemic forecasting based on mobility patterns: an approach and experimental evaluation on COVID-19 data. Social Network Analysis and Mining, 12(1), 1-24. https://doi.org/10.1007/s13278-022-00932-6

[5] Khan, A., Khan, I., Ullah, R., Atangana, A., & Rabiei, A. (2024). Trending on the use of Google mobility data in COVID-19 mathematical models. Advances in Continuous and Discrete Models, 2024(1), Article 38. https://doi.org/10.1186/s13662-024-03816-5

[6] COVID-19 Pandemic Prediction using Time Series Forecasting Models: 2020. https://ieeexplore.ieee.org/abstract/document/9225319

[7] Tsoularis, A., Siettos, C., Anastassopoulou, C., & Russo, L. (2024). Exploring the effects of non-pharmaceutical interventions on COVID-19 transmission through machine learning and modeling approaches. Journal of Mathematical Biology, 89(5), Article 82. https://doi.org/10.1007/s00285-024-02082-z

[8] Zeng, C., Zhang, J., Li, Z., Sun, X., Olatosi, B., Weissman, S. and Li, X. 2021. Spatial-Temporal Relationship Between Population Mobility and COVID-19 Outbreaks in South Carolina: Time Series Forecasting Analysis. Journal of Medical Internet Research. 23, 4 (Mar. 2021), e27045. DOI:https://doi.org/10.2196/27045

[9] COVID-19 Social Mobility: https://socialmobility.covid19dataresources.org/data.html

[10] Gleamviz.org. 2024. GLEAM Project - Global Epidemic and Mobility Model. [online] Available at: https://www.gleamproject.org/

[11] COVID-19 Community Mobility Report: https://www.google.com/covid19/mobility/.

[12] covid-19-data/public/data at master · owid/covid-19-data: https://github.com/owid/covid-19-data/tree/master/public/data.

[13] CSSEGISandData. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://github.com/CSSEGISandData/COVID-19. Accessed: 2024-10-07.