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, vaccination and government mandates into foundational time series models. By acknowledging the significant influence of human movement and social distancing, government Lockdowns, vaccination and other variables 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 all of these variables and Covid-19 infection rates. It will also help us in making much more nuanced predictions.

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 and other exogenous variables in forecasting Covid-19 spread. Moreover, we will integrate comprehensive visualizations 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, vaccination, governmental mandate, 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, SIS 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, vaccination data, and other exogenous variables we aspire to refine these models and enhance their capacity to accurately simulate viral transmission dynamics across various US states.

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, governmental mandates, vaccination policies, and other exogenous variables in influencing viral transmission. External factors such as social behavior, government policies, economic activities, and cultural norms can significantly impact epidemic spread. These dynamics can influence how individuals interact, travel, vaccinate and gather, ultimately affecting the transmission of infectious diseases such as COVID. As factors can vary across regions and time, there is a compelling need to identify and understand the impact of factors with respect to historical epidemiological data to enhance the accuracy of forecasts.

We seek to identify and quantify the relationship between exogenous variables such as mobility, stringency, etc., and COVID infection rates using historical data. Moreover, we plan to evaluate the effectiveness of time-series epidemiological models with and without the integration of exogenous variables in simulating the spread of diseases.

2 UPDATES/CHANGES SINCE PROPOSAL

First of all, we have decided to use three main indicators as exogenous regressors while training the models -

- Mobility Data
- Stringency Index
- Vaccinated Population

Mobility Data has been integrated into the models and used as a regressor in the Seasonal ARIMA model. Stringency Index and Vaccination Population have been integrated into the model and used as the regressors in the Facebook/Meta Prophet model. Both helped increase the accuracy appreciably. Moreover, we've performed interpolation to ensure that the data processing is continuous.

We've utilized two types of mobility information:

- The Mobility Index from the Twitter Mobility Index.
- The Stringency Index from the World Health Organization (WHO) COVID Dashboard.

Most studies utilize only a single kind of mobility information which creates bias. Reliability is improved because multiple data sources have been used. Furthermore, different forecasting models like Seasonal ARIMA & Meta Prophet have been integrated with vaccination data, stringency index, and mobility data.

Exploratory data analysis in section 5.2 and 5.3 emphasizes the fact that mobility data and vaccination data are strong indicators of new COVID cases. Correlations between new cases vs mobility data and new cases vs vaccination data are also explained in

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

a robust manner through plots in sections 5.2 and 5.3. Mobility data has been found to be a lagging indicator and vaccination data has been found to have direct correlation with new cases. We will find similar important metrics to incorporate in all the models to increase model performance.

Moving forward, we aim to fine-tune these models to increase accuracy and implement some more state-of-the-art models. Utilizing multiple parameters such as vaccination, mobility, stringency etc., we plan to achieve higher prediction accuracies. The goal now is to find variables and metrics that directly impact the spread of COVID, and integrate them in our current model implementation and the new models that we will implement. Special emphasis is also given to reliability of resources, and data is collected from multiple sources and interpolated to preserve accuracy.

3 RELEVANT PAST WORK FROM LITERATURE SURVEY

Several studies have aimed to improve epidemic forecasting by integrating social mobility data with time series models. Mobility data from sources such as Google, Apple, Facebook, and Twitter has been utilized in order to provide the most accurate predictions of the COVID-19 spread. Mobility data provides insights into human movement during key phases like lockdowns and re-openings. Utilizing this data, researchers are able to assess how changing mobility patterns impact infection rates.

Research [1] explores the relationship between mobility data and COVID-19 infection rates during the second wave in the United States. Authors considered how public perception and government policies affected virus transmission. The limitations of using tech company mobility data were highlighted such as inaccurate data in less technologically developed regions. The importance of considering factors like mask mandates and precautionary measures, alongside mobility data was also highlighted.

The Twitter Social Mobility Index [2] provided a unique measure of mobility using geolocated tweets. By using this measure, we can understand social distancing and travel patterns across the United States over time. We utilized this paper's approach of using real-time social mobility data.

Mobility network models [3] offer an alternative perspective by focusing on socioeconomic disparities in mobility patterns. According to this study, lower-income areas experienced more volatile infection rates however, the use of network-based approaches over time-series models limited the depth of the analysis conducted. Papers like [4] and [5] highlighted how time-series models, such as LSTM and SIS, could be enhanced with mobility data to predict infection spread. Research paper [4] Demonstrated the effectiveness of LSTM models with mobility data, while [5] emphasized the synergy between Google mobility data and mathematical models like SIS for forecasting infection spread. These studies showcased the potential of combining mobility data and time-series models, and the challenges faced in data complexity and processing.

More advanced time series models were evaluated in [6], where ARIMA and Facebook Prophet were applied to forecast COVID-19 cases. We are utilizing the insights from this study to understand our forecasting accuracy and evaluating our models with metrics such as MAE, RMSE, and MAPE ([6]). Paper [7] incorporated non-pharmaceutical interventions (NPI) such as social distancing and travel restrictions into time series models to enhance their accuracy. The challenge faced was the granular collection of NPI data.

Paper [8] focused on the spatial-temporal relationship between mobility and COVID-19 outbreaks. Poisson count time series models were utilized to demonstrate that mobility was positively correlated with COVID-19 case rates. This study's limitation includes that it used Twitter data exclusively and focused only on a single state. We will use the positive insights from this study in our approach.

Overall, these past works provide crucial insights into the integration of social mobility data with epidemic forecasting models. In our approach, we are utilizing some of the time-series models utilized in the studies above like LSTM, ARIMA, and SI. We are also integrating mobility data to improve our predictive accuracy. The studies above have informed us about the challenges in terms of data complexity, regional variability, and the need to incorporate additional factors such as NPIs. By builing upon these studies, we will further refine the integration of mobility data into time series models to enhance COVID-19 forecasting.

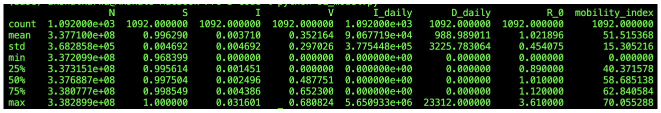
4 DATA COLLECTION PROCESS

The **COVID-19 Data by Our World in Data [12]** dataset that has been collected from the WHO COVID dashboard was utilized in this project. It contains data of shape $(1674, 68)$ for the United States, where 1674 is the number of days and 68 is the number of different features. This dataset was reduced to the 37 most important features. The **COVID-19 Twitter Social Mobility Data [9]** dataset that contains the current index and longitudinal mobility data for several cities, and all the states in the United States was also utilized in this project. Joining this time-series data with the time-series COVID data, we obtained a dataset of 1097 days and 38 features. The first date is January 5, 2020 and the last date is 31st December 2022. In the mobility models Seasonal ARIMA and the Meta Prophet model, the features were further reduced to include only the following information: cases, deaths, vaccinations, date, United States mobility value, and the stringency index. Similarly, in the SIS model with vaccinations and deaths, the following information was used: date, population, infected, vaccinations, United States mobility value.

As we move forward with our project, we will look further into more time series data sources that could complement our project's needs. One such data source is the Github page of the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [13]. Time-series data of COVID-19 cases, deaths and recovery is available here which we will utilize to further enhance our data. Furthermore, our goal is to make all our models even more accurate with more reliable vaccination data, stringency index, mobility data, etc. So, we will be working with other mobility data from that time to further enhance our model with regressive variables.

5 INITIAL FINDINGS & SUMMARY STATISTICS

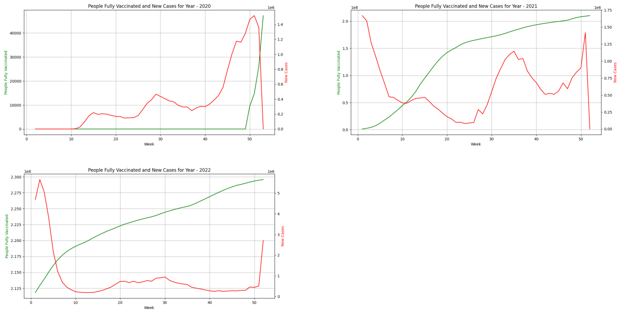
5.1 Preprocessed Dataset Statistics



```
                  N            S            I            V       I_daily     D_daily          R_0  mobility_index
count  1.092000e+03  1092.000000  1092.000000  1092.000000  1.092000e+03  1092.000000  1092.000000     1092.000000
mean   3.377100e+08     0.996290     0.003710     0.352164  9.067719e+04   988.989011     1.021896       51.515368
std    3.682858e+05     0.004692     0.004692     0.297026  3.775448e+05  3225.783064     0.454075       15.305216
min    3.372099e+08     0.968399     0.000000     0.000000  0.000000e+00     0.000000     0.000000        0.000000
25%    3.373151e+08     0.995614     0.001451     0.000000  0.000000e+00     0.000000     0.890000       40.371578
50%    3.376887e+08     0.997504     0.002496     0.487751  0.000000e+00     0.000000     1.010000       58.685138
75%    3.380777e+08     0.998549     0.004386     0.652300  0.000000e+00     0.000000     1.120000       62.840584
max    3.382899e+08     1.000000     0.031601     0.680824  5.650933e+06 23312.000000     3.610000       70.055288
```

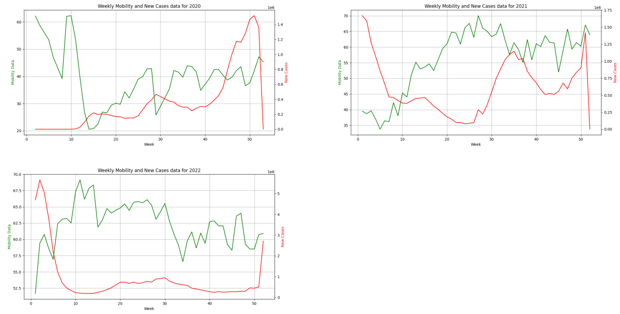
The image shows the summary statistics for the combined dataset:
N: total population size, S: susceptible population size
I: infected population size, V: vaccinated population size
I_daily: people infected daily, D_daily: people dying daily
R_0: reproduction number, mobility_index: mobility value

5.2 Weekly Mobility data vs New Cases for Years - 2020/2021/2022



These graphs have led to the conclusion that mobility data is a lagging indicator of New Covid-19 Cases. Graphs for for New Deaths vs Mobility data and and similar patterns were observed. Therefore, it's extremely important to use factors like mobility data and stringency index as regressors when training the model.

5.3 Vaccination Data vs New Cases for Years - 2020/2021/2022



These graphs have led to the conclusion that vaccination is also a direct indicator of new Covid-19 Cases. As the number of vaccinated people has risen, there have been a low number of new cases that have been depicted. The spikes are due to the waves. But in general, number of vaccinated people is a great indicator and should definitely be used as an exogenous indicator in models.

6 MATHEMATICAL BACKGROUND

The project does not require any advanced or specialized mathematical knowledge that is beyond the fundamentals for this class. Knowledge regarding the following areas is required:

- Data analysis and time-series modeling

- A strong understanding of Python programming and its libraries, such as 'pandas'.

- Data manipulation such as filtering, aggregating, and transforming datasets.

- SIS epidemic model, how infection propagation can be simulated through a population over time based on parameters like the infection rate (β) and recovery rate (γ).

- Time-series forecasting methods such as Seasonal ARIMA (autoregression, differencing, moving averages) and Facebook Prophet.

- Interpolation Methods and Data Standardization/Normalization processes.

- Stringency Index is a composite measure that quantifies the strictness of policies implemented by the government. **Read more about the stringency index.**

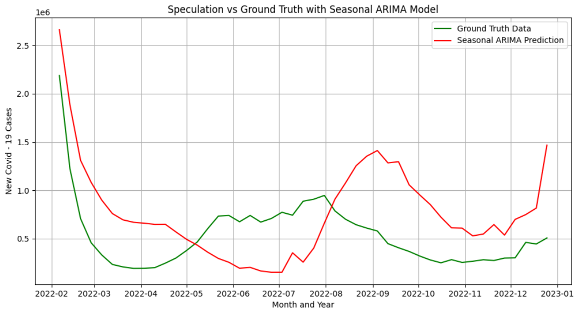
7 ALGORITHMS, TECHNIQUES, AND MODELS

7.1 Seasonal ARIMA (Auto-Regressive Integrated Moving Average)

Seasonal ARIMA is an extension to ARIMA. It's able to incorporate the seasonality of data. Since we had weekly mobility data from the Twitter mobility index, we were able to use this model to do future predictions.

First of all, the data was split into training and test data. The training data was 70% and the test data was 30%. Then, linear interpolation was applied to the mobility data column so as to ensure that the data is continuous and missing values do not impact the accuracy of the model. Then, a seasonal ARIMA model, specfically SARIMAX [14], was used to train the model. The order was kept to be (1,1,1) and the seasonal order was (1,1,1,52) because the data was in weeks. **Mobility Data** was used as the exogenous variable so that model training was done considering Mobility data to be the external regressor.

After the model training and fitting, we forecasted the predictions on the testing data against the ground truth labels. The results were very promising. The model was accurately able to predict the trend in the rise and fall of Covid-19 cases.



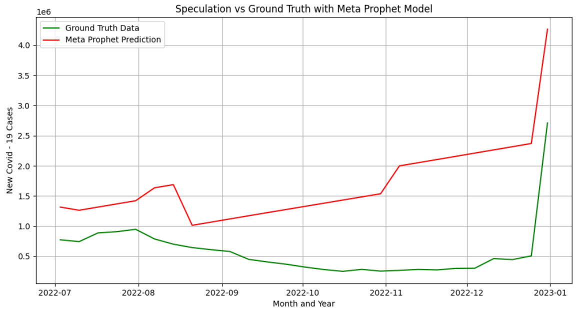
Seasonal ARIMA Prediction with Mobility Data as the regressor. Moving forward, the model will be further fine-tuned with other exogenous regressor variables like vaccination and stringency index. Also, more fine tuning will be done Auto-regressive, Integrated and Moving Averages order to achieve even better accuracies. That being said, the current model is fairly accurate and is able to model spread quite accurately.

7.2 Facebook Prophet

The Open Source model - Facebook Prophet - was used to forecast disease spread efficiently. Some initial data scoping was needed for Facebook Prophet to be trained because of the unique way the model requires the inputs to be. 'ds' and 'y' columns are needed so as to train the data. The **ds** column can be considered to the x variable and the **y** column can be considered to the target or y variable.

Then, 2 years and 6 months of data is used as the training data and 6 months as the testing data. The Facebook Prophet model is initialized with weekly seasonality turned on. The model is trained with two regressors - **Fully Vaccinated Population** and **Stringency Index**. Then, the model is trained and fitted so as to make predictions.

This model was also quite strong in depicting the trends - rise and fall of covid-19 cases.



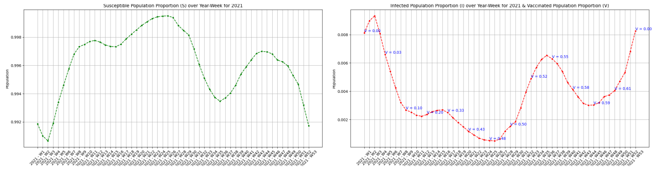
Facebook Prophet Prediction with Stringency Index and Total Vaccinated Population as the regressors

Moving forward, more fine-tuning will be done with additional regressors and parameters so as to achieve higher accuracy.

7.3 SIS Model with Vaccinations & Death

The SIS (Susceptible, Infected, Susceptible) model with vaccinations and death was implemented. Even if a person recovers from COVID, they can still be susceptible again and be a carrier as well. The COVID vaccine reduces the viral load of individuals thereby reducing the chances of getting infected. This does not mean that the individual is recovered and can never get infected again.

The SIS model that we have implemented updates the overall and the infected population size by accounting for deaths. Based on a recovery time parameter, the susceptible and infected population sizes are updated as well.



The plots show the susceptible (green) and the infected (red) population proportions vs weeks in 2021. In the plot on the right, we see the vaccination population proportion in blue. Initially, we can see that we increasing vaccinations the infected population reduces. Then, we can see an increase which was due to a new variant followed by another local minima with increasing vaccinations.

Moving forward, we plan to enhance the model by conducting spatio-temporal analysis using various parameters.

7.4 Hidden Markov Models (HMM)

The overarching goal is to also implement a HMM model which would be utilized to predict the spread of a virus based on historical epidemic data. The plan is to similarly use different regressors here with Covid-19 data to accurately predict the spread of the virus.

7.5 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. Similar training and test data split will be done as in ARIMA and Facebook Prophet. We plan to implement some learnings and approaches from papers [4] and [5]. Something that we are also considering is that if all our models have considerable accuracy. We can create an ensemble model with weighted averages of these models with their own different regressors. This will help us in accurately predicting the spread of the virus.

8 DIFFICULTIES & CHALLENGES

One of the major difficulty challenges that we faced was combining the data from both the sources to use for training and evaluation purposes. To determine which population index to use since we were doing week wise analysis and one of the data sources has day wise data for vaccination and stringency index. Then creating an amalgamation of data-frames to run the analysis in an efficient manner. Secondly, standardizing the data format was a difficult task. While shrinking data, it was difficult to standardize a process for which day should we be choosing. But, we realized using the cumulative metrics for each week and using the last date of the week is the most efficient way of solving the problem. There were also some date formatting intricacies, so that index can be created. However, we were able to resolve them by further understanding how model source has been written and adapting our data for that use case. Fine-tuning the models was interesting and challenging. We tried with different regressors and parameters to check for model performance and accuracy. In this way, we were able to tune the models and adapt for the ones giving best accuracy.

In the SIS model with vaccinations and death, it was important to calculate the infected population proportion accurately. In case of COVID, all recovered individuals become susceptible again. Determining the number of days in which they become susceptible again and accurately utilizing this parameter was a challenge. For now, I have 14 days as the model parameter because this is approximately the maximum number of days. Realizing the accurate value of this parameter (between 8-14 days) would improve our model results and we plan to further research this. Accounting for different external factors in the SIS model is another difficulty that we aim to overcome.

9 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 aim to compare the performance of different time-series models, both with and without incorporating variables such as mobility data. This will help us assess the impact of these variables on model accuracy and predictive capabilities.

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 variables like mobility in 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.

10 PROJECT GOALS AND IMPACT

10.1 Goals

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

Quantify the Relationship between variables such as mobility and COVID-19 spread: By analyzing COVID-19 time-series data and external factor variables, the objective is to uncover correlations between these factors and infection trends to identify key indicators that influence virus transmission.

Evaluate and Enhance Forecasting Models: In addition to SIS with vaccination and deaths, time-series models like seasonal ARIMA, and Facebook Prophet, we will implement and compare several other diverse models such as Hidden Markov Models and LSTM. Comparing them, we will evaluate how accurately each tracks and predicts COVID-19 spread with and without the integration of the other parameters such as mobility. Lastly, we will also refine these models to enhance their accuracy and predictive capabilities.

Visualize the Impact of Factors like Mobility, Stringency on Epidemic Dynamics: We will develop comprehensive visualizations, such as plots for each model with and without each parameter, to illustrate how different parameters influence infection rates across different U.S. states, major cities and regions.

Assess 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.

10.2 Impact

With the integration of variables such as mobility, stringency index, etc., our project improves the accuracy of current state of the art models by accounting for multiple external factors. By understanding the influence of external factors, its easier to mitigate the impact of future outbreaks. Our project thus bridges the gap between epidemiological models and real-world social dynamics.

10.3 Stretch Goals

In addition to mobility and stringency index, we aim to narrow the gap further by accounting for features such as smoking history, diabetes prevalence, climate factors, economic status, population density, and more, to enhance our model's predictions further. We aspire to contribute to the development of more reliable and responsive epidemic forecasting tools.

11 TIMELINE

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

(1) **Phase 1 (11/4 - 11/8)**: Improving dataset quality by researching and joining 2-3 more datasets in our master dataset such as the [13].
(2) **Phase 2 (11/8 - 11/15)**: Implementing 2-3 more models such as LSTM and HMM with baseline parameters and exogenous regressors. Expanding on our SIS model to implement regressor variables. More versions of time series SI models like SIERV, SEIHR, SIS, etc.
(3) **Phase 2 (11/15 - 11/20)**: Hyper-parameter tuning and integration of variables such as mobility, stringency index, etc.
(4) **Phase 3 (11/20 - 11/25)**: Evaluation of model performance across diverse metrics.
(5) **Phase 4 (11/25 - 11/29)**: Creating all the visualizations for each model and each different variable's impact on each model.
(6) **Phase 5 (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.

[14] Seabold, S. and Perktold, J. statsmodels: Econometric and statistical modeling with Python. statsmodels 0.12.2 documentation. https://www.statsmodels.org/devel/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html. Accessed: 2024-11-04.