Bridging the Gap: Enhancing COVID-19 Epidemic Forecasting by Integrating Factors like Vaccination Rates, Mobility, Stringency, Socio-Economic Indicators, and Health Metrics 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 accuracy of COVID-19 epidemic forecasting by integrating factors such as new vaccinations, mobility data, stringency index, human development index, diabetes prevalence, and GDP per capita into core time series models. Taking into account the significant impact of vaccination rates, social mobility, government mandates, and socioeconomic factors on viral transmission, an analysis of historical COVID-19 data is conducted in conjunction with these factors. This approach allows us to explore the correlations between these factors and infection rates leading to more nuanced and reliable predictions.

Our methodology involves implementing a variety of time series forecasting models to simulate the spread of COVID-19, to evaluate their ability to capture the influence of social mobility and other exogenous factors on epidemic dynamics. Comprehensive visualizations are utilized to illustrate the predictions of different time series forecasting models in comparison to COVID-19 ground truth data. The effectiveness of state-of-the-art epidemiological models in replicating observed COVID-19 transmission is assessed.

Models such as SARIMA, SI, Facebook Prophet, TBATS, RNN, and LSTM are thoroughly analyzed and their results are contrasted with actual COVID-19 trajectories to identify both their strengths and weaknesses. Many models also incorporate data on social mobility, vaccination rates, and other exogenous variables, these models are refined and their ability to accurately simulate viral transmission dynamics across different U.S. states is improved.

2 Response to Milestone Comment

Building on the feedback for the Project Milestone, the following changes have been made:

- The set of exogenous variables has been expanded to include metrics like New Vaccinations, Human Development Index, Diabetes Prevalence, GDP Per Capita, Mobility Data, Stringency Index, etc. These additional metrics provide us with a comprehensive picture of the influence of external factors during the spread of COVID.
- Advanced deep learning models based on Recurrent Neural Networks: LSTM and RNN, have been added. The TBATS model (**T**rigonometric seasonality **B**ox-Cox transformation **A**RMA errors **T**rend **S**easonal components) has also been added. The predictions of these models have been analyzed.
- Analysis has been performed with unique sets of exogenous variables. Using the exogenous variables in isolation as well as in diverse combinations, performance of different models has been analyzed. This gives us a deep interpretibilty analysis for each variable for each specific model and also how strong each variable's lag and correlation for each model is.
- Graphs have been added for some variables that show how each one of them would indicate the correlation and lag of the spread of the virus.
- After taking different sets of exogenous variables, performing model analysis, and looking at forecasting accuracies, the best combination of exogenous variables is found.
- In-depth analysis has been performed on how each exogenous variable performs in isolation as well as in combination with others and the graphs and discussion has been added.
- "New Vaccinations" and "People Fully Vaccinated" have been introduced as regressor variables to account for new vaccinations and understand the changes in disease spread trends before and after vaccinations. These variables have been used in isolation and have also been used with other regressor variables to understand how different models forecast in presence of different external factors. This helps account for the changes pre and post vaccinations and allows us to better understand the impact of other variables.
- For example, Figure 2, (b) People Fully Vaccinated and Stringency Index have been used as exogenous variables to make the predictions. This takes into account the time when vaccination hasn't started and also when vaccination started. Another example, is in Figure 1, (d). In this way, by using New Vaccinations and People Fully Vaccinated as regressor variables, the effects of variables before and after vaccinations started in making better predictions have been analyzed.
- The contributions of each author have been added in Section 8 - Author Contributions.

3 Problem Statement

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

,
2024.

mandates, vaccination policies, and other exogenous variables like GDP 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, vaccinations, human development, diabetes, GDP per capita, 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.

4 Relevant Past Work & 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.

One 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 SI, 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 SI 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 and many other exogenous factors like vaccination, GDP, human development index 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 Non Pharmaceutical Interventions (NPIs).

5 Proposed Method

5.1 Intuition

Predicting pandemic spread remains one of the most complex challenges in epidemiological modeling. The traditional approaches that exist often provide inaccurate results because they treat disease transmission purely as a mathematical phenomenon. They rely on simple differential equations or basic time series models that fail to account for the complicated relationships that exist. There are many real-world factors to consider that influence modern epidemics. This project suggests a better approach by recognizing a critical insight: ***pandemic spread is a multidimensional phenomenon deeply embedded in social, economic, and behavioral ecosystems***.

The project is state-of-the-art as it integrates a variety of important exogenous variables to accurately capture the complex dynamics of the spread of COVID:

- Human Development Index (HDI): Provides insights into a population's overall socioeconomic resilience.
- GDP Per Capita: Reflects the healthcare accessibility and the economic capacity of a region to respond to crises.
- Diabetes Prevalence: Helps consider the population's health vulnerability.
- Vaccinations: Allows to keep track of the real-time immunization progress.
- Social Mobility Index: Captures the populations' movement patterns.

- Stringency Index: Measures the effectiveness of strategies of containment.

Overall, by utilizing a diverse set of exogenous variables, the forecasting ability of models is enhanced. The models were tested with and without different combinations of exogenous variables and their performances were then compared. The models being used are: LSTM, RNN, Meta Prophet, SI model with vaccination and deaths, TBATS, and SARIMA.

5.1.1 Why is this State-of-the-Art?

Problems with Conventional Epidemiological Forecasting Models:

- Models that are prevalent these days depend on a very small group of variables and parameters like initial infection rate, density of the population, etc.
- Conventional models operate under the preemptive concepts which are not able to capture the nuances of epidemic spread in different social, geographical, and resource-based areas.
- One of the major drawbacks of the models is that the population is considered to be homogeneous. For this reason, the diverse variations in healthcare resource availability and infrastructure, economic conditions, human development, vaccination capabilities are not infused into the calculations properly.

For this project, we've taken a robust and comprehensive approach so as to ensure that we are able to change how epidemic trend forecasting is done. Extremely crucial and rudimentary sets of exogenous variables with different permutations and combinations are integrated so as to ensure that the extremely complex association between the epidemic spread and regressor/exogenous variables is captured valuably.

New Methodology: There are multiple methods and approaches that are used in this paper. All of them add an extra layer to the existing one-dimensional model prediction strategy. Advanced deep learning models like RNN and LSTM are used. Other models like Meta Prophet, SI, and SARIMA are also use with different set of exogenous variables to maker sure that the prediction trends are cohesive. The predictions are then made with i) no regressor variables, ii) regressor variable iii) different permutations and combinations of regressor variables. These predictions are the matched with the test/ground data that we have. This helps us in making sure that the different performance metrics with different sets of variables is studied effectively to see which leads to the best performance.

How is this helpful? The robust models that have the context for the entire set of regressor variables - vaccination, human development index, GDP, etc. lead to a much better outline which can be given to policy makers in the government and public servants who make health related decisions. This will in-turn help in making sure that the right resources are used at the right place at the right time. If we can make sure that the interventions are effective, then epidemics can be curbed in a much more efficient manner. Also, something else to note is that predictions for epidemics is an extremely stochastic process, we shouldn't look at it like a deterministic process that *could* would give us the *right* answer.

The **Main Difference** is that conventional models are juxtaposed with regressor variables which in-turn allow to alleviate epidemic forecasting into a multi-dimensional problem with different parameters like GDP, vaccinations, mobility, etc.

The methods and processes that we have explored in this paper build a layer on top of the existing epidemiology methods. We have to make sure that epidemiological models are extremely context aware, compatible, and flexible with a lot of external factors and metrics. This notion is explored in this paper.

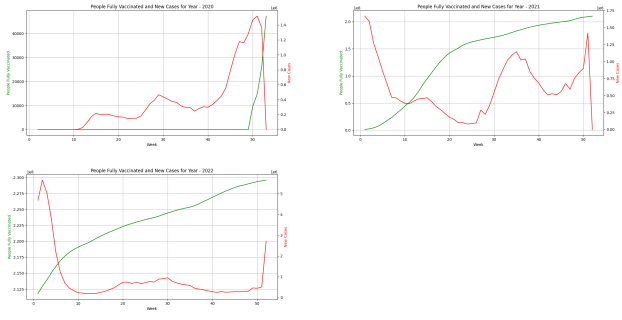
5.2 Description

5.2.1 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, Stringency Index, People Fully Vaccinated, Human Development Index, Diabetes Prevalence, and GDP Per Capita. Similarly, in SI model with vaccinations and deaths, the following information was used: date, population, infected, vaccinations, United States mobility value. Then, in LSTM and RNN models the same dataset has been used but predictions have been made without exogenous variables. The goal for both of those models was to adapt them to time series Covid-19 data and make accurate predictions.

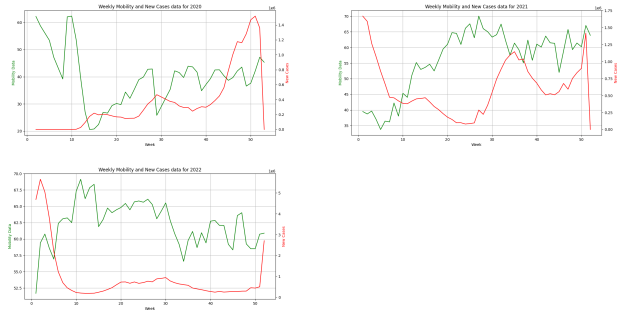
5.2.2 Weekly Mobility data vs New Cases for Years - 2020/2021/2022



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.

5.2.3 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.

Other exogenous variables like Diabetes Prevalence, GDP Per Capita, Stringency Index, etc. are also directly or indirectly related to the spread of the virus. However, to prevent duplicacy we've not included all the graphs in the paper.

5.2.4 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 have weekly mobility data from the Twitter mobility index, we were able to use this model to do future predictions. The ability for SARIMA to do time series forecasting across different seasonal patterns is excellent. Also, the ability to juxtapose them with the exogenous variables while model training is also quite impressive. This makes it a very strong fit for our use case.

Four sets of exogenous variables were used in-order to tune to the predictions. After the model training and fitting for each of these cases, predictions are forecasted on the testing data against the ground truth values and performance is measured.

5.2.5 Facebook Prophet

The Open Source model - Facebook Prophet - was used to forecast disease spread efficiently. The Facebook Prophet model can also be used with data that has strong seasonal patterns. Also, regressor variables can be incorporated during model training and prediction itself. This makes the model a great fit for our use case.

Some initial data cleaning 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.

Four sets of exogenous variables were used in-order to tune to the predictions. After the model training and fitting for each of these cases, predictions are forecasted on the testing data against the ground truth values and performance is measured.

5.2.6 SI Model with Vaccinations & Death

The SI (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 SI model 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.

5.2.7 Recurrent Neural Network (RNN) Models

A Recurrent Neural Network (RNN) based approach was utilized to predict the spread of COVID-19. Learnings and approaches from papers [4] and [5] were implemented.

- **Single-Layer RNN Model:** A simple recurrent neural network with a single layer of 100 units and 0.2 dropout rate was utilized to capture short-term temporal dependencies in COVID-19 spread data through a basic sequential learning approach.
- **Two-Layer RNN Model:** An expanded RNN architecture featuring two SimpleRNN layers (100 units each) followed by two Dense layers (10 units) was utilized to capture more intricate temporal patterns in epidemic spread data.
- **Single-Layer LSTM Model:** A single LSTM layer with 100 units designed to learn long-term dependencies in COVID-19 transmission sequences was utilized. LSTM's advanced memory mechanisms allowed for more nuanced temporal pattern recognition.
- **Two-Layer LSTM Model:** A multi-layer LSTM architecture combining two LSTM layers (100 units each) with two Dense layers (10 units) was utilized to get enhanced predictions where sequential data was processed through multiple sophisticated memory layers.

5.2.8 TBATS model

TBATS (Trigonometric, Box-Cox, ARMA, Trend, and Seasonal) is also a Time series forecasting model that can handle complex time series data that has diverse seasonality. For our use case, the TBATS model is very informative. The model is also very flexible. We used this model in-order to see if it's able to comprehend the seasonality in our time series data better than SARIMAX. It was indeed able to do so. The model is very powerful and gives accurate patterns for the spread without even using exogenous variables. It's used by scientists when SARIMA/ARIMA fails, and we wanted to use this model to try if it could lead to even better predictions. We've predicted data for the last 7 weeks and interpolated the new cases column in case the the value is 0 or data is missing.

6 Experiments/Results

6.1 Experimental Questions & Testbed

This project aims to evaluate the performance of various epidemic forecasting models in predicting the spread of COVID-19. The effectiveness of traditional models like the SI model (with vaccinations and deaths) and advanced techniques such as SARIMA, Facebook Prophet, LSTM, RNN, and TBATS are investigated. The primary questions addressed include:

- How accurately do these models forecast COVID-19 spread?
- What role do external factors like mobility index, stringency index, new vaccinations, GDP Per Capita, Human Development Index, etc. play in improving model predictions?
- How well each model captures short-term and long-term trends, and handles factors like seasonality?
- How do vaccination rates and deaths impact epidemic dynamics?

The testbed for this experiment uses real-world COVID-19 time-series data. Different models have different ranges from the ground truth based on model implementation and external factors. External indicators such mobility, stringency, vaccination, GDP, human development, etc. are also integrated to assess their impact on forecasting accuracy. The dataset covers multiple regions in the US and time periods. Models are implemented using state-of-the-art libraries, with hyper-parameter tuning to optimize performance.

6.2 Experimental Analysis

6.2.1 Seasonal ARIMA (Auto-Regressive Integrated Moving Average)

Training data is 70% and the test data was 30%. Linear interpolation is applied to the regressor variables. Then, a seasonal ARIMA model, specfically SARIMAX [14], is used to train the model. The order was kept to be (1,1,1) and the seasonal order was (1,1,1,52).

Four sets of exogenous variables were used in-order to tune to the predictions. a) Mobility Data, b) Mobility Data, Stringency, c) Mobility Data, Stringency Index, GDP Per Capita, d) Mobility Data, GDP Per Capita, New Vaccinations.

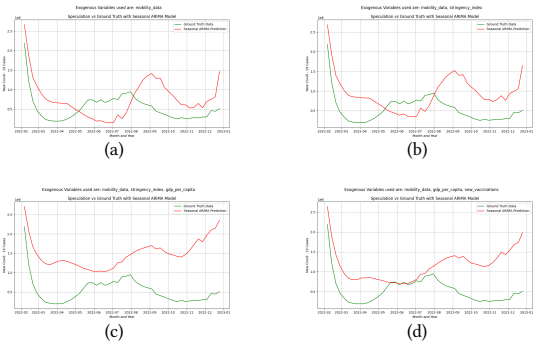


Figure 1: Predictions for SARIMAX using different sets of regressor variables

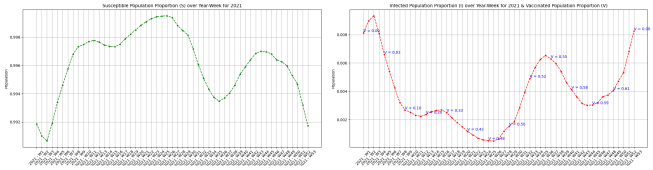
The results are quite impressive. We can see that the best prediction trends are actually in subplot (d) where we have used the three regressor variables: Mobility Data, GDP Per Capita, and New Vaccinations. The subplot (b) has stringency index along with mobility data. But the predictions are almost the same as compared to (a). This is because both the variables indicate almost the same thing. Also, for (c) we've added GDP per capita as an exogenous variable and the results are a little worse. This combination doesn't lead to a better prediction. However, the best combination that predicts the most accurate results is (d) with Mobility Data, GDP Per Capita, and New Vaccinations. Due to this we can say that vaccination is an extremely crucial metric. As vaccination increased, cases reduced and it was an excellent variable for prediction. In the code, the model can be run for any combination of exogenous variables. Such models are great to predict the trend of the spread of the virus and are excellent to use in-order to curb future spread.

If we use only one exogenous variable in the SARIMAX model, the performance of each exog variable is as follows based on RMSE values:

New Vaccinations > Human Development Index > Diabetes Prevalence > GDP Per Capita > Mobility Data > Stringency Index

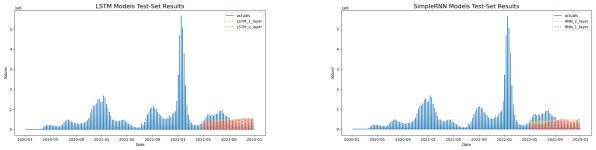
Also, something to note here is that models like these are great to measure the general trend of cases as you can also see from the graphs.

6.2.2 SI Model with Vaccinations & Death



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, it is visible that with increasing vaccinations the infected population reduces. Then, an increase can be witnessed, which was due to a new variant, followed by another local minima with increasing vaccinations.

6.2.3 Recurrent Neural Networks



The **LSTM** and **SimpleRNN** model results shown in the two images above illustrate how well these neural network architectures capture the overall trends in COVID-19 spread. The LSTM models demonstrate a strong ability to track the general trajectory of the actual COVID-19 values over time. They may not have perfectly predicted the sharp peaks, however, they were able to closely follow the broader rise in new cases. Similarly, the SimpleRNN models also provided with a good prediction on the general trend of COVID-19

spread. They captured the underlying patterns in the time series data. RNN models are extremely good at generating accurate predictions. The models' ability to learn temporal dependencies makes them valuable predictive instruments. Accounting for other external factors while predicting with RNNs can help understand some of the sudden spikes observed in the data. Overall, RNNs made great predictions on the test set of about 8 months of data.

6.2.4 Facebook Prophet

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. Then, the model is trained and fitted so as to make predictions.

For this model also, 4 sets of regressor variables were added: a) Human Development Index, Mobility Data b) People Fully Vaccinated, Stringency Index, c) Stringency Index, GDP Per Capita, d) Mobility Data, People Fully Vaccinated, Stringency Index, Human Development Index, Diabetes Prevalence, and GDP Per Capita.

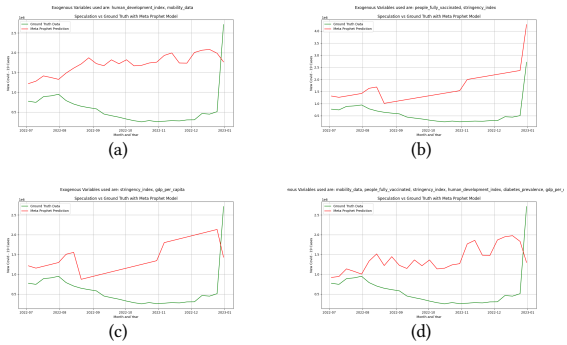


Figure 2: Predictions for Facebook Prophet using different sets of regressor variables

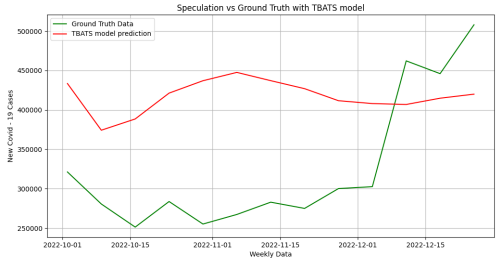
In this case, the best results are observed in (b) when only 2 regressor variables are used: People Fully Vaccinated and Stringency Index. This means that the model is able to follow the terminology that when *vaccination is not prevalent, then cases are much higher* and vice versa. Also, when a lot of regressor variables are used as in (d), then the predictions also start to hallucinate a little bit and are not that accurate. However, the model performance gets better and better when as seen in (a) and (c) as compared to the base model (without exog variables) when exogenous variables are added. Overall, Facebook prophet is very strong in depicting the trends rise and fall of Covid-19 cases. For different exogenous variables, graphs can be seen and the model can also be run for different sets of these variables from the code. That being said, the results for each of the 4 cases are very intuitive and interesting in understanding how predictions are different when different exogenous variables are used.

If we use only one exogenous variable in the Facebook Prophet model, the performance of each exog variable is as follows based on RMSE values:

GDP Per Capita > Diabetes Prevalence > Human Development Index > New Vaccinations > Stringency Index > Mobility Data

Also, something to note here is that models like these are great to measure the general trend of cases as you can also see from the graphs.

6.2.5 TBATS model



TBATS is the model which is able to comprehend seasonal complexities in time series data. It's able to comprehend seasonability and make predictions accurately. The model performance is very powerful so as to give us accurate trends on a weekly level. The entire data set is used to predict the values and trends for the last 7 weeks. The new cases data column is also interpolated in-case of any missing data. That being said, it was very informative to implement a new and powerful model and do weekly forecasting and get accurate prediction trends.

6.3 Findings

Each model is very unique. The data is so diverse that it's almost impossible to predict the exact values. Metrics like RMSE are not that relevant here because we're trying to predict the trend of virus spread. Trends have been predicted very aptly and accurately. We have also mentioned the order in which exogenous variables should be used in-order to know which variables are the most relevant.

For the SARIMAX model, we've done analysis for which combination of regressor variables could be the best in order to predict trends. We've also done analysis on singular exogenous variables and how they impact the models performance.

For the Facebook Prophet model also, we've done analysis for which combination of regressor variables could be the best in order to predict trends. We've also done analysis on singular exogenous variables and how they impact the models performance.

As you can see via the metrics, the exogenous variables are extremely helpful in helping navigate accurate predictions. As the number of different exogenous variables is increased, the prediction in general gets more and more accurate. Different exogenous variables used individually in different models behave differently.

The TBATS model doesn't have a regressor but it's able to capture the weekly trends extremely accurately. We decided on using this because we wanted to test a seasonal model's performance on weekly prediction without the use of a regressor. The graph accursedly depicts the prediction trend that it has been able to give.

The RNN_2_Layer model outperformed the others, achieving the lowest RMSE (95,198.32) and competitive MAE (36,336.57). The LSTM_1_Layer model had slightly higher RMSE (98,318.51) but

the lowest MAE (34,102.96). RNN_1_Layer and LSTM_2_Layer performed significantly worse, with higher RMSE and MAE values.

Observations: 1092 (1)

Dynamically Tested: *True* (2)

Test Set Length: 245 (3)

CI-Level: 0.95 (4)

The above values correspond to all the four RNN models. The model with the best results was: **Best Model:** RNN_2_Layer. The evaluations results for the four RNN models are in Table 1.

Table 1: Performance of Different Models

Model Name	Test Set RMSE	Test Set MAE
RNN_2_layer	95,198.32	36,336.57
LSTM_1_layer	98,318.51	34,102.96
RNN_1_layer	133,052.04	47,039.23
LSTM_2_layer	216,635.11	77,186.12

The **RNN_2_Layer** model emerged as the most effective, consistently delivering accurate predictions with minimal errors. The **LSTM_1_Layer** model also demonstrated strong performance in reducing absolute error. RNN_1_Layer and LSTM_2_Layer underperformed and showed higher error rates. The results suggest that its important to select appropriately balanced architectures for time-series forecasting tasks.

7 Conclusion & Discussion

The project tests the predictive capabilities of state-of-the-art time series forecasting models by using COVID pandemic data. Advanced models such as SARIMA, Meta Prophet, SI with vaccination and deaths, LSTM, RNN, and TBATS were implemented and the significant potential of having an adaptive system approach in epidemic forecasting rather than a fixed model approach was demonstrated. The complex relationships between COVID- 19 case predictions and external variables such as mobility, health parameters, and socio-economic factors (Vaccination rates, Stringency index, Mobility index, GDP per capita, Human Development Index) were explored. Through the utilization of a diverse set of models including traditional methods and advanced deep learning approaches, the project demonstrates the challenges of accurately predicting new cases. Predicting the trajectory of new cases is however, extremely valuable. Forecasting new cases can provide significant insights into how the government and health organizations need to handle situations to limit the spread of any disease.

The findings highlight that certain variables such as vaccination rates and stringency index serve as highly effective regressor variables. These, alongside mobility data and the diabetes prevalence variable play a critical role in determining the dynamics of the epidemic in regions. They should therefore be prioritized when predicting disease spread. It is also critically important to choose the correct combination of exogenous variables. A combination of exogenous variables along with accurate weights for each can certainly improve forecasting. Different stages of pandemics can have varying significance of exogenous variables and this highlights the growing need for dynamic and adaptive modeling strategies. For instance, the mobility index might have played a more crucial role during initial lockdown periods, while economic indicators like GDP per capita would have had greater predictive power during economic recovery phases. This suggests that developing a weighting mechanism for exogenous variables can substantially enhance prediction accuracy.

There is significant potential for future research into more dynamic and adaptive models that allow for weighted exogenous variable inclusion. Moreover, these models should allow time-series analysis for external factors that can continually keep changing. Models that dynamically adjust variable weights based on temporal context can truly transform epidemic forecasting.

This project lays the foundation to build upon when developing epidemic forecasting models that account for exogenous variables. Future work can focus on creating ensemble models that combine the strengths of several approaches. Incorporating additional sets of exogenous variables and fine-tuning existing models can also enhance prediction accuracies. Furthermore, the development of better epidemiological models can help better understand the role of exogenous factors in disease spread across different regions and populations.

8 Author Contributions

All the work has been done by both the individuals - Mehul Rastogi and Akshat Karwa. Equal contributions were made to the design and implementation of the project. The data cleaning process was conducted together by both the individuals. Mehul implemented the Meta Prophet model, the SARIMA model and the TBATS model. Akshat implemented the SI model, the RNN and LSTM models. Both participated in the brainstorming, testing, debugging, and refining of the project. Through close teamwork and shared responsibilities, the authors collectively shaped this project from conceptualization to state-of-the-art for epidemic forecasting.

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.