**Literature Review**

Our team reviewed several articles regarding forecasting models to project the spread of the covid-19 pandemic. From the outset in March 2020, researchers, data analysts, and machine learning engineers applied various time series models from Simple Exponential Smoothing to ARIMA to Deep Learning Neural Networks to Facebook’s Prophet all in an attempt to alert public health officials about the spread of the virus and predict a possible end date. As stated in the article, “COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population” by Vasilis Papastefanopoulos , Pantelis Linardatos and Sotiris Kotsiantis[3], the authors state: “Being able to accurately forecast when the outbreak will hit its peak would significantly diminish the impact of the disease, as it would allow governments to alter their policy accordingly and plan ahead for the preventive steps needed such as public health messaging, raising awareness of citizens and increasing the capacity of the health system.”

Our review, which also included covid-19 time series projects on Kaggle, discovered the following similarities across all of the materials:

*Data Collection*

Data from Johns Hopkins University (JHU) and the World Health Organization (WHO) were the most common source of information for these articles. The data covered the time period from the early days of the pandemic to the summer of 2020. With the benefit of hindsight, our model will be able to expand on the time frame of data collection as well as have access to more complete data. It is an established fact that there were major problems with data collection in the early days of the pandemic. The covid tracking project at The Atlantic magazine created a scoring system for how states reported their data.[[1]](#footnote-1) Since March 7, 2021, The Atlantic covid tracking project has stopped tracking data due to the fact that the CDC has improved its data collection and reporting.

*Models*

Several different mathematical and time series models were used to forecast the infection rates. Ping-En Lu Yi-Cheng Chen and Cheng-Shang Chang’s paper, “A Time-Dependent SIR Model for Covid-19 with Undetectable Infected Persons”[11], used the time-dependent model, susceptible-infected-removed (“SIR”) that tracks transmission and recovery rates. Ian Cooper,, Argha Mondal, and Chris G. Antonopoulos also applied the SIR model in their paper, “A SIR model assumption for the spread of COVID-19 in different communities “[12]. They “augmented the classic SIR model with the ability to accommodate surges in the number of susceptible individuals, supplemented by recorded data from China, South Korea, India, Australia, USA, Italy and the state of Texas in the USA to provide insights into the spread of COVID-19 in communities. In all cases, the model predictions could be fitted to the published data reasonably well, with some fits better than others.”

The most common time series approach is exemplified in the work by Vasilis Papastefanopoulos, Pantelis Linardatos, and Sotiris Kotsiantis[3]. They used six different time series approaches, Auto-Regressive Integrated Moving Average (“ARIMA”), the Holt–Winters additive model (“HWAAS”), TBAT (an acronym for the four fundamental components of the method: Trigonometric seasonal formulation, Box–Cox transformation, ARMA errors and trend component), Facebook’s Prophet, DeepAR (a forecasting method based on autoregressive recurrent neural networks for probabilistic forecasting), and N-Beats (Neural Basis Expansion Analysis for Interpretable Time Series Forecasting). The models were applied across 10 nations. Using RMSE to assess the models, they found “traditional statistical methods such as such ARIMA and TBAT overall prevail over deep learning counterparts such as DeepAR and N-BEATS—an outcome which, due to the lack of large amounts of data.”

Our models will improve on the work that we have reviewed so far given that the quality of data collection has improved, and we have more information to tune the various time series models to make our predictions.

**Key Words**

*Pandemic, Covid, Time Series, MAPE, RMSE, ARIMA, Prophet, Cases, Predictions, SARIMA, Forecasting, Comparison, Long short-term Memory (LSTM), Recurrent Neural Network (RNN), Bidirectional LSTM (BiLSTM), Gated Recurrent Units (GRUs), and Variational AutoEncoder (VAE), Support Vector Machines (SVM)*

**References**

1. **Kumar, Naresh & Susan, Seba, “COVID-19 Pandemic Prediction using Time Series Forecasting Models”, *The 11th ICCCNT 2020 conference*.** In this paper, two specific time series models were picked to see which one would perform better when trying to forecast the spread of the covid 19 virus. The two models used to model the prediction of the covid 19 pandemic was a ARIMA model and the model from Facebook called Prophet. The dataset used for both models are from the github repository maintained by John Hopkins University. The primary means of evaluation in determining which model had better performance was the criteria of MAE, RSME and MAPE. It was concluded by the research team that ARIMA did a better job of predicting future covid related cases than the Facebook Prophet model.
2. **Er, Başak, Emeç, Murat, and Ozcanhan, Mehmet, “Analysis of COVID-19 Data Using Arima Time Series Model”, *Conference: V. International Scientific And Vocational Studies Congress – Engineering* (December 2020).** In this journal paper, instead of comparing different models to determine which time series model could best predict the covid 19 pandemic outbreak. The researchers decided they wanted to know if one specific time series model could help to make predictions about the spread of covid 19 and be used to make decisions from. And that specific model was the ARIMA model. They also used the same data held in the github repository maintained by the renowned John Hopkins University to study the ARIMA models prediction of the Covid 19 virus spread. It was concluded that the ARIMA time series forecasting model can be effective in prediction Covid 19 related case but its effectiveness is in very short-term forecasting. Short-term as in less than a month. Anything over a month, the model had much higher errors.
3. **Papastefanopoulos, Vasilis, Linardatos, Pantelis, and Kotsiantis, Sotiris, “COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population” *Applied Sciences. 10. 3880. 10.3390/app10113880* (May 2020)**was reviewed. In this journal paper, the researchers took a similar approach to that of another set of researchers that wanted to compare several time series forecasting models and see which one would have the best outcome in predicting covid related cases caused by the covid 19 pandemic. This time a total of 5 different models were chosen and used to model covid 19 predictions. The 5 models were ARIMA , Holt–Winters additive model ,TBAT , Facebook’s Prophet, Deep AR, and N-Beats. The N-Beats model is one of a neural network model that for the previous comparison of models did not include. The researchers’ primary measurement in determining better accuracy between the models was RSME. It was determined at the end based on the RSME values that TBATS and ARIMA which the researchers made a point to point out were very traditional and not new models were two best models and prevailed over more newer and perhaps what come people may consider more flashier ones.
4. **Mahmud, Sakib, Bangladesh COVID-19 Daily Cases Time Series Analysis using Facebook Prophet Model, *Social Science Research Network*** was reviewed. For this paper, a study was done to predict daily cases of covid 19 using only the Facebook Prophet model. The data collected for this time series prediction model was for Bangladesh only with a date range of July 22nd 2020 to September 19th 2020. The source of the data was taken from a website called Our World in Data. This study came to almost the same conclusion as a previous study had when it came to assessing the results of the predictions of Covid 19 cases from the use of the Facebook Prophet model. And that conclusion was that for daily predictions and predicting only for very near term future, the model was able to perform much better than for predictions where the horizon was much larger. And that was mostly due to the model expecting the trend to keep growing and rising over time over the long term. A previous that used more classical models such as ARIMA and exponential smoothing had better results than Facebook prophet model which is considered a newer model that was introduced more recently.
5. **Ismail, Khan, Znati, Materwala, Turaev, ”Tailoring time series models for forecasting coronavirus spread: Case studies of 187 countries”, *Computational and Structural Biotechnology Journal Volume 18, 2020, Pages 2972-3206* (September 2020)**. In this paper, a study was conducted just like the previous studies on comparing the different time series models and their predictions future Covid 19 cases based on historical time series data collected at a pre defined time range. The participants of this study used similar performance measures to compare the results objectively and that was RSME and MAPE. All the models they used will not be listed but they used only the classical models such as ARIMA and very simple ones like the moving average model. Their results were also similar to one of another group that had also incorporated these models to do direct comparisons. And ARIMA performed the best when it came exponential increasing trend whereas for trend that was both linear and exponential , the Holts Linear Trend model outperformed the field. One difference in this study was that it studied over 187 different countries. And because of that, the researchers concluded that what country the model was based on and the trend that each country exhibited makes a difference in the performance of the models. In essence the model needs to match the country and has to be specific to the country its being modeled on and cannot just be used as a one size fits models on all countries to achieve the models’ best results.
6. **Kumar,Vinay, Chimmula, Reddy, and Zhang ,Lei “Time series forecasting of COVID-19 transmission in Canada using LSTM networks****”, *Chaos, Solitons and Fractals Nonlinear Science, and Nonequilibrium and Complex Phenomena* (May 2020)**. Using data from Johns Hopkins University and the Canadian Health authority, the authors developed a forecasting model using the Long short-term Memory (LSTM) network, a deep learning approach. The authors were motivated to use a non-linear approach for different levels of the outbreak and other changes such as seasonal variations. Given that the outbreak is a time series, the authors recommend using sequential networks to extract the patterns from it. Earlier attempts to use LSTM networks were not able to represent the spatio-terminal components simultaneously. The authors overcame this by modifying the internal connections between the input and output cells. The models were trained with data collected until March 31, 2020 as reported by Canadian health authority. Results were measured using the MSE. Additionally, the authors created a second LSTM model using data trained on the Italian dataset to predict Canadian cases. The authors concluded that the data revealed that prompt public health safety measures taken by the Canadian government had positive impact on transmission rates as compared to the U.S. and Italy.
7. **Chaurasia, Vikas and Pal,Saurabh, “Application of machine learning time series analysis for prediction COVID-19 pandemic”, *Sociedade Brasileira de Engenharia Biomedica* (October 2020)**. In this study, the authors used data taken from the WHO’s “Data WHO Coronavirus Covid-19 cases and deaths-WHO-COVID-19-global-data” from the period January 22, 2020, to May 28, 2020 to implement several forecasting techniques (naive method, simple average, moving average, single exponential smoothing, Holt linear trend method, Holt-Winters method and ARIMA) with the goal of trying to minimize the RMSE. Their results showed that the naïve method had the lowest RMSE. The authors made this point about the data: “The data is unstable; it also shows that the number of deaths has increased exponentially since mid-March 2020. Another issue facing the study is insufficient training data. Four months of (January 2020 to April 2020) data are used for training purposes, 29 days of verification data, based on which the number of deaths can be determined expected in the coming months. There are very few training data for machine learning to train itself. Moreover, the number of infected people changes rapidly; the case occurred in mid-March.”
8. **Zeroual, Abdelhafid, Harrouc ,Fouzi, Dairi, Abdelkader, and Sunc, Ying, “Deep learning methods for forecasting COVID-19 time-Series data: a Comparative study”, *Chaos, Solitons and Fractals Nonlinear Science, and Nonequilibrium and Complex Phenomena* (July 15, 2020**) this study compared five deep learning methods to forecast the number of new cases and recovered cases. The five networks are Recurrent Neural Network (RNN), Long short-term memory (L STM), Bidirectional L STM (BiL STM), Gated recurrent units (GRUs) and Variational Auto Encoder (VAE). The data was collected from Italy, Spain, France, China, USA, and Australia. These five models were chosen due to their ability to handle temporal dependencies in time series data, distribution-free learning models, and their flexibility in modeling nonlinear features. The models were evaluated on data collected from Johns Hopkins from the start of the pandemic to June 17, 2020. Their key finding was that the VAE model outperformed the other models. “The efficiency of the VAE model for COVID-19 forecasting is promising and manifested. This fact is maybe due to the capacity of the VAE in dealing with small data compared to the other recurrent models (RNN, L STM, Bi-L STM, and GRU) which may need more lengthy data to extract relevant variability in time series data.” “The VAE can capture almost all variability in data and provide more accurate forecasting in comparison to the other RNN-based models. All other models perform moderate forecasting performance in terms of RMSE, MAE, MAPE, and RMSLE and show poor performance in terms of explained variance. This is maybe due to their need for more data in the training to capture the dynamics of COVID-19.”
9. **Vijander Singh , Ramesh Chandra Poonia , Sandeep Kumar, Pranav Dass , Pankaj Agarwal , Vaibhav Bhatnagar & Linesh Raja, “Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine”, *Journal of Discrete Mathematical Sciences and Cryptography*, 23:8, 1583-1597, DOI: 1080/09720529.2020.1784535.** Using worldwide data collected from Johns Hopkins University Public Repository Center for Systems Science and Engineering (CSSE) for the period Jan. 22, 2020 to April 25, 2020, the authors created a model to predict the confirmed cases of the COVID-19 using Support Vector Machine (SVM) which is unique among other models. SVM models are widely used for classification algorithms. They found that SVM produces optimized performance values to forecast the predicted COVID-19 cases using kernel function. “Support Vector model is suitable for time series dataset and we have applied SVM model to predict the confirm, deaths and recovered patients. The SVM kernel is a tuning parameter that takes input space of low dimensions and transforms it into a higher dimensional space, i.e. translates nonseparable problem into a separable problem. Here ‘poly’, ‘sigmoid’ and ‘rbf’ has been used as the problem is linear problem.”
10. **Shah, Saloni; Mulahuwaish, Aos; Ghafoor, Kayhan; Maghdid, Halgurd S., “Prediction of Global Spread of Covid-19 Pandemic: A Review and Research Challenges.” *TechRxiv. Preprint*.** [***https://doi.org/10.36227/techrxiv.12824378.v1***](https://doi.org/10.36227/techrxiv.12824378.v1) In this paper, the authors present a comparison of day level forecasting models on COVID-19 affected cases using time series models and mathematical formulation. Their forecasting models “…strongly suggest that the number of coronavirus cases grows exponentially in countries that do not mandate quarantines, restrictions on travel and public gatherings, and closing of schools, universities, and workplaces (“Social Distancing”).” Using data from Johns Hopkins and the WHO, the authors created simple time series forecasting models using Moving Average (MA), Weighted Average (WMA), and Single Exponential Smoothing (SES). These models are very basic time series models. MA depends on the assumption that future observations are related to an average of recent observations. WMA modifies MA by assigning weights to different observations. Finally, SES is a smoothing time series data based on the exponential window function. Each model was judged using Mean Absolute Deviation (MAD), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). They found that SES is the most accurate model for forecasting confirmed cases, recovered cases, and deaths.
11. **Yi-Cheng Chen, Ping-En Lu , Cheng-Shang Chang, “A Time-Dependent SIR Model for COVID-19 with Undetectable Infected Persons”, *IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, VOL. 7, NO. 4*, (OCTOBER-DECEMBER 2020)** In this paper, the authors conduct mathematical and numerical analyses for COVID-19. To predict the trend of COVID-19, we propose a time-dependent SIR model that tracks the transmission and recovering rate at time t. Using the data provided by China authority, we show our one-day prediction errors are almost less than 3%.
12. **Ian Cooper, Argha Mondal , Chris G. Antonopoulos, “A SIR model assumption for the spread of COVID-19 in different communities”, *Chaos, Solitons and Fractals 139* (2020)** The authors the effectiveness of the modelling approach on the pandemic due to the spreading of the novel COVID-19 disease and develop a susceptible-infected-removed (SIR) model that provides a theoretical framework to investigate its spread within a community. Here, the model is based upon the well-known susceptible-infected-removed (SIR) model with the difference that a total population is not defined or kept constant per se and the number of susceptible individuals does not decline monotonically.

1. https://covidtracking.com/analysis-updates/weve-launched-a-new-state-grading-system [↑](#footnote-ref-1)