CUNY DATA 698 Master's Research Project Spring 2021
Mid Term Draft Report

A Comparison of Covid-19 Time Series Models Forecasting Cases, Hospitalizations, and Deaths for New York

Santosh Cheruku
Angel Claudio
John K. Hancock
John Suh
Subhalaxmi Rout

Introduction

In December 2019, the World Health Organization ("WHO") reported the earliest onset of symptoms of a new highly contagious disease in Wuhan, China. By January 2020, the WHO confirmed human to human transmission of this new virus. This disease which is caused by the virus Severe Acute Respiratory Syndrome Coronavirus 2 or SARS-CoV-2 became known globally as COVID-19. The virus is spread through aerosol transmission of human respiratory droplets.

By March of 2020 the WHO declared the virus a pandemic. The United States became the epicenter of the pandemic by the end of March 2020 with New York reporting the highest number of cases in the world with over 12,000 cases reported in early April 2020. The exponential rise of cases created a strain on hospital resources, and the number of deaths also grew at an alarming rate.

The severity and rapid transmission of COVID-19 caught government officials off guard. There were shortages of Personal Protection Equipment ("PPE") for frontline health care workers, hospitals were over run, and shipping containers had to be used for the over flow of the deceased. The crisis could have been mitigated with better data collection and time series models that would have informed government officials of the potential impact.

The objective of this project is to compare multiple time series models, from the naïve to the more advanced, to determine which model can accurately predict cases, hospitalizations, and deaths from this pandemic. The best model(s) could be used by public health officials to prepare for future pandemics.

Can a time series model created from data taken from the early months of the pandemic provide accurate forecasts for a future period in 2020? As an example, could time series data taken from March through May of 2020 predict cases, hospitalizations, and deaths in June 2020? Could that same model provide a reliable forecast for July 2020?

Literature Review

Our research found an abundance of articles that used time series models to forecast the spread of covid-19. We saw that from the start, in March 2020, when the World Health Organization ("WHO") declared COVID-19 a global pandemic, researchers deployed an array of time series models to forecast the spread of the virus. These efforts were largely done to inform the general public of the need to enact mitigation measures to stop the spread. The articles discussed in this review provide an exemplar of the articles that we researched. We begin our review by comparing the findings from very basic time series models and finishing with a look at the more advanced ones.

Authors Haytham H. Elmousalami and Aboul Ella Hassanien's research focused on the day level spread of the virus [1]. Their methodology used naive time series models, Moving Average ("MA"), Weighted Moving Average ("WMA"), which take averages or weighted averages of past observations to forecast future cases. Such models are easy to create and even easier to report to the public. Most media outlets include a moving average component when reporting about the spread. The other model that the authors used was Single Exponential Smoothing ("SES"), another naïve model. In this instance, instead of weighing past observations equally, SES uses functions to exponentially decrease the weighting of past observations.
The results of Elmousalami and Hassanien's models showed that the SES model had the highest accuracy for confirmed cases, recovered cases, and deaths based on the evaluations of Mean Absolute Deviation ("MAD"), Mean Square Error ("MSE), Root Mean Square Error ("RMSE"), and Mean Absolute Percentage Error ("MAPE"). In contrast to Elmousalami and Hassanien, authors Vasilis Papastefanopoulos, Pantelis Linardatos and Sotiris Kotsiantis used a higher class of time series models [2], Auto Regressive Integrated Moving Average ("ARIMA"), Holt-Winters additive model ("HWAM"), Trigonometric seasonal formulation Box-Cox transformation ARIMA errors and trend component ("TBAT"), Facebook's Prophet, Deep AR, a probabilistic forecasting with Auto-Regressive Recurrent Networks, and N-Beats, a neural basis expansion analysis for interpretable time series forecasting.

Papastefanopoulos, Linardatos, and Kotsiantis' models consisted of linear regression and deep learning neural networks. Instead of applying averages or weights or exponentially decreasing weights, these models primarily make predictions by using either a regression of past observations or by using a system of inter-connected nodes that learns from past observations. Also, in contrast to Elmousalami and Hassanien, the authors Papastefanopoulos, Linardatos, and Kotsiantis did not find a "one-size-fits-all" model. Their findings showed that based, on RMSE measures, the ARIMA and TBAT models performed best in most of the countries while achieving second best in the other two. 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 [2]." (emphasis added)

Authors Vinay Kumar, Reddy Chimmula, and Lei Zhang focused solely on using a single deep learning network, Long short term Memory ("LSTM") [3], a non-linear approach that uses a Recurrent Neural Network ("RNN") to forecast trends. In an RNN, output from the last step is fed as input to the current step. This is somewhat similar to Elmousalami and Hassanien's approach where the weights applied to past observations are manipulated to make forecasts. The difference is that in LSTM networks can retain long term information which is useful if there are lags of unknown duration between important time gaps.

Using data collected in Canada until March 31, 2020, Kumar, Chimmula, and Zhang's methodology was to use sequential networks to extract the patterns from a time series dataset. The rationale for this approach was that the linear approach often neglects the temporal components in the data. "They depend upon regression without non- linear functions and failed to capture the dynamics of transmission of infectious diseases like novel corona virus. Statistical models such as Auto Regressive Integrated Moving Average (ARIMA), Moving Average (MA), Auto Regressive (AR) methods overwhelmingly depends on assumptions and such models are difficult for forecasting real-time transmission rates." [3] In contrast to Papastefanopoulos, Linardatos, and Kotsiantis, they showed that the RMSE of the LSTM had the highest accuracy.

Authors Abdelhafid Zeroual et al. [4] differed from the previously discussed studies by comparing the five most advanced models, Recurrent Neural Network ("RNN"), Long short-term memory ("LSTM"), Bi-directional LSTM ("Bi-LSTM"), Gated recurrent units ("GRUs") and Variational AutoEncoder ("VAE") to forecast cases and recovered cases across six countries. The authors cited the models' ability to "handling temporal dependencies in time series data, distribution-free learning models, and their flexibility in modeling nonlinear features." [4] Using RMSE as their primary performance metric, the authors found that VAE outperformed the other models for confirmed and recovered cases. This study was one of the first times that VAE has been used to model COVID-19 cases. The authors offer a reason as to why VAE outperformed the other advanced models. "[T]he 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 [4]." (emphasis added) Poor performance of advanced models due to the lack of data was the same issue that Papastefanopoulos, Linardatos, and Kotsiantise experienced with the performance of DeepAR and N-BEATS. However, we did not see this issue with Kumar, Chimmula, and Zhang's LSTM model.

The articles discussed and compared in our review are indicative of the articles that we researched for the project. Each one chose a particular class of models to compare, used various datasets, compared the results using standard metrics, and summarized their findings. What we did not find was one paper that compared all classes of models, from the naive to the more advanced, in one study over one single dataset over the same period of time in order to determine which model(s) performed the best. That is the objective of this paper.

Key Words

Pandemic, COVID, Time Series, MAPE, RMSE, ARIMA, Prophet, SARIMA, Forecasting, Long short-term Memory ("LSTM"), Recurrent Neural Network ("RNN"), Bidirectional LSTM ("BiLSTM"), Gated Recurrent Units ("GRUs"), and Variational AutoEncoder ("VAE")

Methodology

Overview

The research problem that this project attempts to solve is to find the most efficacious time series model that could have provided health officials with the most accurate predictions of the number of COVID-19 cases, hospitalizations, and deaths in the early days of the pandemic. To make this determination, we will compare the results of multiple time series models ranging from the very basic to the most advanced.

To start, we reviewed data collected by authoritative sources, the Centers for Disease Control ("CDC") and Johns Hopkins University, for the time period from March to June 2020. After review of these and other sources, we made a selection of the best data source for analysis. We then proceed to conduct an Exploratory Data Analysis and prepare the data for the models by looking for missing values, outliers, and any other anomalies.

Once we have a prepared data set, we will build 12 time series of various classes, from the most basic, Moving Average, Weighted Moving Average, Simple Exponential Smoothing, Holt Winters, to the next level, ARIMA, Auto Regressive, to the more advanced, Facebook Prophet, XGBOOST, Neural Network, and LSTM, . Each model will be written in Python, and will be evaluated using standard metrics, Root Mean Square Error ("RMSE") and Mean Absolute Percentage Error ("MAPE"). The performance of each model on the same dataset over the same time period and the same geographical location should provide a solution to our research problem.

Data Sources

Our team investigated three data sets as candidates for modeling future effects of COVID-19 in areas resembling the state of New York. The first of the three was the Centers for Disease Control and Prevention (CDC) [15]. The CDC data, while extensive, did not have any attributes for location, which was vital for us to create a model for our main area of interest (New York) and then trying to contrast it with a similar location (e.g., Florida, Texas, California).

The next data source we vetted was the Johns Hopkins University Center for Systems Science and Engineering [16]. The data set was a strong contender for modeling, but we found the volume of the data (424 features) did not come without the need for a lot of data manipulation. This included filtering out hundreds of unnecessary columns, a pivot execution in

both directions for values and columns needing restructure, multiple data cleansing tasks including imputations, and finally we would still be left with a heavy data source for extracting and using for computations.

This finally left us with our selected data set, The COVID Tracking Project at the Atlantic [17]. The data was readily available for us to extract via a web API and since it is a compilation of data brought together by the Atlantic team of data scientists and analysts, the data for the most part was clean and manageable. Inclusive of this we included data to explore public transportation ridership from the Metropolitan Transportation Authority [18] in New York. To explore additional effects of behavior we also added an attribute depicting what dates in our time series fell on dates that were a national holiday [19] (or dates of known seasonal gatherings such as spring break which garners large crowds of people, increasing the risk of spreading a virus.

Our platform for this project is on Python version 3.8. The main libraries used for the data wrangling portion is "pandas", "io", and "matplotlib" with "seaborn" for creating EDA extrapolations and visualizations. IDE choice varied per team member based on preference - for example PyCharm by Jet Brains and Spyder open-source software were used during development. Microsoft's GitHub was used as version control repository for project documents, resources, and executable code.

The COVID Tracking Project at the Atlantic supplies a public web API for data consumption. We were interested in region specific data sets and fortunately had the option to supply the query string of the API URL with any choice name of a US state. Unfortunately, the hosting of this resource had a schedule to be taken down on May 1, 2021. To make our code robust to this future take down, on every data pull we backed up the data to a physical file and included a "Try-Catch" block to use that physical file in the event the API raised an error or simply did not return any data.

The data itself had a manageable structure and clean data. There were many features excluded as a result of not being applicable such as "date modified" or having one value for every observation - for example, a variable with exclusively "NA" values. In conclusion, we considered half of the 56 original columns and joined two columns to the time series, a binary feature indicating time of holiday, and a numeric value to indicate public transportation ridership from the MTA of New York.

Our first secondary data of holiday data was corroborated online and by common knowledge into a hard coded list. We then looped through the time series and if a timestamp fell in that list of holiday times a binary value was inserted to indicate such. The next secondary data, New York's MTA ridership time series, was able to be joined simply by using the date index. In conclusion after every run, we back up the final data set to a physical CSV file as a disaster recovery method in lieu of our various data sources being corrupted or absent.

Technology
Programming Language: Python (version 3.8)
IDE: Jupyter Notebook, PyCharm, Spyder
Libraries: Pandas, Numpy, matplotlib, seaborn, io, warnings, datetime
Collaboration: Github
Communication: Slack, Zoom, Google meet

References

1. Haytham H. Elmousalami and Aboul Ella Hassanien, “Day Level Forecasting for Coronavirus Disease (COVID-19) Spread: Analysis, Modeling and Recommendations”, Scientific Research Group in Egypt (SRGE), Cairo, Egypt (March 15, 2020)

2. 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)

3. 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).

4. 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)

5. Kumar, Naresh & Susan, Seba, “COVID-19 Pandemic Prediction using Time Series Forecasting Models”, The 11th ICCCNT 2020 conference.

6. 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).

7. Mahmud, Sakib, Bangladesh COVID-19 Daily Cases Time Series Analysis using Facebook Prophet Model, Social Science Research Network

8. 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).

9. Chaurasia, Vikas and Pal,Saurabh, “Application of machine learning time series analysis for prediction COVID-19 pandemic”, Sociedade Brasileira de Engenharia Biomedica (October 2020).

10. 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.

11. 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

12. 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)

13. 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)

14. Vinay Kumar, Reddy Chimmula, and Lei Zhan, “Time series forecasting of COVID-19 transmission in Canada using LSTM networks”, Chaos, Solitons and Fractals (May 2020)

15. Centers for Disease Control and Prevention, ‘United States COVID-19 Cases and Deaths by State over Time’, 2021. [Online]. Available: https://data.cdc.gov

16. Johns Hopkins University, ‘Coronavirus Resource Center’, 2021. [Online]. Available: https://coronavirus.jhu.edu/

17. The Atlantic, ‘The Covid Tracking Project’, 2021. [Online]. Available: https://covidtracking.com

18. Metropolitan Transportation Authority of New York, NY, ‘Day-by-day ridership numbers’, 2021. [Online]. Available: https://new.mta.info/coronavirus/ridership

19. U.S. Office of Personnel Management, ‘Policy, Data, Oversight’, 2021. [Online]. Available: https://www.opm.gov/policy-data-oversight/pay-leave/federal-holidays/