Departmemt of Mathematics and Statistics

CAPSTONE PAPER REVIEW WEEK 1 (PAPER 1)

Time Series Analysis on COVID-19: Literature Review on Prediction and Analysis of COVID-19 Daily New Cases and Cumulative Cases: Time Series Forecasting and Machine Learning Models

Gianyce Michelle Gesualdo Ortiz

Supervisor: **Dr. Shushen Pu**

Contents

CHAPTER 1

Background/Motivation

The article "Prediction and Analysis of COVID-19 Daily New Cases and Cumulative Cases: Time Series Forecasting and Machine Learning Models" is occurring during the end of 2021 and attempts to predict December 1, 2021, to December 30, 2021. The epidemic of COVID-19 became prevalent in 2019 for the first time as a type of respiratory disease, threatening global public life and security. The authors chose the USA, India, and Brazil to observe because "as of December 28, 2021, the cumulative prevalence of coronal pneumonia was most outstanding in [those] countries".

The problem this specific article addresses involves the different models used in the past to predict COVID-19 prevalence and mortality rates in recent studies. The authors explore ways to forecast COVID-19 spread using machine learning and statistical techniques. They argue that the best prediction results are not achieved with a single tool, noting that a standalone time series analysis model is not enough to capture the nonlinear aspects of the disease's progression. Also, they need to find ways to capture the behavior of an epidemic in the medium and in the long term.

As an individual who lived through the effects of the pandemic, exploring these types of epidemics and ways to predict their forecasting can aid in public health initiatives by creating aid and resources for under-insured populations. Additionally, awareness of seasonal spikes and potential outbreaks allows the public to prepare appropriately and enables companies to provide adequate supplies, offering protection against the inevitable. This article highlights that these models explore the predicted effects of time series and machine learning methods on the future of an epidemic to find

risks in advance, avoid outbreaks, and use data for decision-making. This article also touches on other articles that explain how the implications of time series analysis have changed the future of analytics.

CHAPTER 2

Methods Used

The authors utilize different models to analyze and forecast the COVID-19 data, each chosen for their specific reason. One of the primary models selected was the Automatic Regressive Integrated Moving Average (ARIMA) model. It has some advantages in its simple structure and immediate applicability. This model has been previously used to model different diseases as well. This model is also capable of correlating regulation with short-term changing trends in the time series. Its effectiveness in capturing short-term fluctuations makes it reliable for analyzing time-sensitive data like COVID-19 cases. This paper primarily used the ARIMA model for predicting cumulative cases, showing that it performs best with data with positive growth trends.

The second mentioned model is Facebook's Prophet Model. Itâs an open-source model that can handle time-series data with the advantages of taking strong seasonal effects, missing data, outliers, and trend changes. This makes it well-suited for the dynamic and often unpredictable nature of COVID-19 data, where trends can shift rapidly due to various external factors. This model was applied to predict daily and cumulative new cases, especially with irregularities in the data. This model also demonstrated strong performance in handling data with more significant fluctuations and seasonal patterns compared to the other models. Specifically, this model could be seen performing the most accurately within the USA.

Then, the SARIMA model, which is an extension of ARIMA, is often used for time series forecasting as it accounts for seasonality. Seasonality in data refers to the regular and predictable changes that occur. With diseases in general, seasonality is common.

For instance, flu season is commonly expected every year. This model extrapolates the state of something at some point in the future by analyzing the pattern of historical data based on that pattern from the past to make future predictions. Within this specific paper, the SARIMA model also showed a strong performance in handling data within a shorter, seven-day period in daily new cases of COVID-19.

The decision to use these models was based on the nature of the COVID-19 data, which is characterized by its random, nonlinear, and seasonal nature. The authors considered these factors when selecting the models. After extracting data from the WHO website, they cleaned and processed the data to remove unnecessary information and ensure it was stationary. With the prepared data, they ran the ARIMA, SARIMA, and Prophet models to conduct their analysis and make predictions. These models were then ran using R 4.1.1 software, and they used the forecast and prophet packages. The authors used RMSE, MAE, and MAPE to analyze the data by accuracy and effectiveness, which demonstrated the different accuracies that each of the models had in predicting COVID-19.

The paper focuses on the possibilities and capabilities of forecasting COVID-19 and the ability to identify future trends with these model adaptations. After being processed through R software, each of these methods has different capabilities for this research. For instance, the ARIMA model as a statistical analysis tool is powerful at analyzing and forecasting time-series data. In this specific case, it was beneficial for the research question as it has an ability to model positive growth trends effectively. It simulates and estimates the state of something at some point in the future. Notably, when the P, D, and Q values in ARIMA are set to 0, it changes the model significantly. When $q = 0$, it works as an A.R. model, when $p = 0$, it becomes the M.A. model, so both the (p, q) are important factors to determine the ARIMA model.

Parameters used in the ARIMA model were as follows: p, which is the autoregressive order; d, which is the degree of differencing; and q, the moving average order. The optimization process comes from the outcome of ACF, autocorrelation function, and PACF, partial autocorrelation function analysis. The ACF helps in identifying the q parameter, and PACF helps in determining the p parameter. The ACF describes the degree of correlation between the current value of that sequence and its past value. The PACF is similar, but rather than finding correlations of lags like ACF with the

current value, it finds correlations of the residuals with the next lag value. The ACF helps identify the q parameter, and the PACF helps determine the p parameter. These concepts helped construct the ARIMA and SARIMA models, as they were analyzed for the smooth time series.

After testing the combinations of p, d, and q, they utilized the Bayesian Information Criterion (BIC) to help identify optimization. BIC is a class of information criteria to measure the goodness of fit for a statistical model. It builds on the concept of entropy, which can be considered a measure of how well a model captures the underlying data patterns without overfitting. Meaning, it can weigh the complexity of the estimated model against the goodness of fit of this model of the data. These qualities help assess the model's parameters and how it performs. Within this context, the model with the lowest BIC is preferred as it indicates a better balance between model fit and complexity. All three, BIC, ACF, and PACF, were used in model construction and training. The optimal combination of the three was identified and performed.

For accuracy evaluation, they used coefficient of determination and compared their training model to the SARIMA model. This showed that the R^2 of the SARIMA model was between 0. And 1.0, meaning that the prediction accuracy of the prophet model is higher and can be applied to the actual prediction of COVID-19. Also, in this particular study, the SARIMA model showed signs of overfitting with poor generalization ability, whereas the prophet model was fitted the cases and captured the seasonality hidden. The Prophet model effectively predicted the daily new COVID-19 cases in the USA, as seen by the lower RMSE and MAE values compared to the SARIMA model.

Prophet Model's Results:

RMSE: 13,437.603

MAE: 7,118.961

SARIMA Model's Results:

RMSE: 14,850.734

MAE: 7,877.085

Comparing the RMSE and the MAE values of both, it is evident that the Prophet model has the lower of the two. A lower RMSE indicates that, on average, the predictions made by the prohet model were closer to the actual observed values. The lower MAE indicates that the absolute differences between the predicted values and the actual

values were smaller, therefore indicating the modelâs predictions were more accurate.

The study also revealed that ARIMA model performed better in Brazil and India, demonstrating that for data with fluctuation/seasonality, Prophet is well suited. Still, for forecasting cumulative cases with consistent growth, ARIMA works better.

In other words, the ARIMA model can help identify how certain regulations or interventions might influence or correlate with the immediate, short-term trends in the data. For example, if a new policy is implemented to control the spread of a disease, the ARIMA model could potentially show how this policy correlates with a decline or change in the number of cases in the short term.

CHAPTER 3

Significance of the Work

This paper showed that comparing the three models, ARIMA, SARIMA, and Prophet, successfully predicted the daily and cumulative cases of COVID-19 in the USA, Brazil, and India. Depending on the type of model sheds light on what findings were discovered. As mentioned previously the Prophet model was found to be effective in predicting daily new positive growth trends in Brazil and India. ARIMA showed an ability to highlight a strong performance in predicting cumulative cases in Brazil and India. Overall, the study did show us how to handle Big Data for time series analysis in a way that applied multiple models that can be tailored in different ways to show us that data with seasonality, nonlinearity, and missing values can still be analyzed proficiently.

The paper's overall significance also highlights the importance of selecting appropriate models for analysis, the future implications of these models, and how to use data analytics in public health applications.

Aside from epidemiology, these models and analyses can be applied to many different nonlinear data types that show seasonality. For example, rainfall, electrical currents and stock market prices. If we can continue to find models to create more accurate forecasting, it is just the beginning to the power behind these models. Also, combining these techniques with Random Forests or neural networks to improve the accuracy and robustness is the future of these types of technologies.

CHAPTER 4

Connection to Other Work

The references refer to an important work entitled âTime Series Analysis: Forecasting and Controlâ by George E. P. Box and Gwilym M. Jenkins. This seminal work is the foundation of time series analysis, the use of ARIMA, and how it is implemented to forecast time series data. This work laid the groundwork for applying time series models.

The introduction of this paper discusses the challenges that previous models faced in capturing the full complexity of the COVID-19 pandemic. As the epidemic spread rapidly, scientists struggled to accurately model its seasonality and progression. Anastassopoulou et al.'s paper, "Data-based analysis, modeling and forecasting of the COVID-19 outbreak," published at the peak of the pandemic, was crucial for providing a standard against which this paper builds to enhance the modeling and forecasting of COVID-19. Although perhaps not the most vital work, it allowed for papers like this one to come forth in a way that opened doors for many other time series types. Especially since this paper did not do anything with time series or with integrating prophet modeling, which is not often seen with older epidemiology papers.

CHAPTER 5

Relevance to Capstone Projects

My group's paper focuses on creating a time series analysis of COVID-19. Continuing to learn and understand methodologies that I could potentially apply in my capstone project, including SARIMA, the Prophet model, and ARIMA has allowed me to widen my current understanding of time series analytics. However, the project conducted by Wang, Yanding, et al. was more in-depth than the project my group is conducting. Additionally, I draw heavily on my interest in bio informatics and statistics to guide the direction of my research. Exploring the field of epidemiology is also crucial for me as a data scientist, as it enables me to apply the various machine learning and statistical techniques that I have learned during my Master of Science program. This integration of methodologies and my personal interests provides a comprehensive approach to understanding and forecasting the trends associated with the pandemic.

Bibliography

[1] Wang, Yanding, et al. "Prediction and Analysis of COVID-19 Daily New Cases and Cumulative Cases: Time Series Forecasting and Machine Learning Models. *BMC Infectious Diseases, vol. 22*, 2022

[2] Box, George E. P., and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control Holden-Day*, 1970