Departmemt of Mathematics and Statistics

CAPSTONE PAPER REVIEW WEEK 2 (PAPER 1)

Literature Review on Predicting the New Cases of Coronavirus in India using Time Series Analysis as Machine Learning Model in Python

Gianyce Michelle Gesualdo Ortiz

Supervisor: **Dr. Shushen Pu**

Contents

1 Background/Motivation 3

2 Methods Used 4

3 Significance of the Work 6

4 Connection to Other Work 7

5 Relevance to Capstone Projects 8

CHAPTER 1

Background/Motivation

This article was written in August 2020 during the early COVID-19 epidemic. It highlights the severity of COVID in India specifically and aims to use models like ARIMA and the AR model to predict new cases in India. When the paper was written, very little was known about COVID-19. The impact was suddenly severe to the world as a public threat. The paper addresses that the Government of India, after thousands of deaths, issued lockdown multiple times to try and reduce the death toll. Knowing this, the literature presents data and visualization of India's spread using Python programming. The research aims to enforce the use of computational studies to strengthen public help and preventative measures toward battling certain epidemiological cases within our society. It is not necessarily specific to COVID-19 since the learned techniques can be used to model and forecast any pandemic/epidemic case in the future.

CHAPTER 2

Methods Used

The paper addresses time series data, represented as a series of observations over particular time periods, as well as cross-sectional and pooled data. Cross-sectional data refers to data from one or more variables collected at the same time, while pooled data combines time series and cross-sectional data.

ARIMA models, powerful yet simpler autoregressive moving averages that add the notion of integration, are used to analyze and forecast time series data. ARIMA models are denoted with the following:

p: Number of lag observations, i.e. lag order.

d:Number of times raw observations are differenced.

q: Size of the moving average window, i.e. order of moving average.

ARIMA models, being a subset of linear regression models, possess a unique adaptability. A value of 0 in any ARIMA parameter can configure the model to perform different functions, which the authors use to determine if ARIMA or AR would be better suited for the prediction.

Data was collected and downloaded from Jan.1, 2020, to July 31, 2020. The data was run with the autocorrelation function, and it was noted that due to the behavior of the lag, the graph can only have a maximum of 4. The author then creates a training set from the new cases in India. They chose an 80-20 split, where 80 is the percentage from the total data set, set, and the remaining 20 comes from the test dataset. From the ARIMA model, the $(2, 2, 4)$ combination was selected because, at those values, the AIC was at its lowest, which suggested that it had the best fit with minimal complexity

relative to the other available combinations. Using that, they forecasted step = 17 for the predicted ARIMA model.

The AR model was similarly implemented as the ARIMA, with the exception that $p = 0$. From the plotted figures, it could be seen that the AR model would require more fine-tuning.

These methods were used specifically to see the predicted future using time series analysis on COVID-19 in India. The methods were well suited for the data and the research methods since ARIMA models are commonly used for this type of analysis. However, it would have been better for the author to use a SARIMA model, as the epidemic trends more seasonally and rapidly.

CHAPTER 3

Significance of the Work

Although this work is more basic compared to other papers I have read on using ARIMA for COVID-19 analysis, its early release during the pandemic has made it a frequently referenced and inspirational piece for many subsequent time series analysis studies on COVID-19. This paper also highlights why using only one model for prediction may not be the most reliable method to be used for a time series analysis, as the ARIMA is subject to biases if the residuals are not close to zero, as non-zero residuals indicate that the model has not fully captured the patterns in the data.

CHAPTER 4

Connection to Other Work

My group reviewed a recent work, "Prediction and Analysis of COVID-19 Daily New Cases and Cumulative Cases: Time Series Forecasting and Machine Learning Models." This paper refers to Kulshretha's paper as a way to expand beyond the ARIMA model. It talks about the importance of using more models to capture the data, as biases and model strength can change and be more accurate with other model types.

CHAPTER 5

Relevance to Capstone Projects

My group's literature reviews are focused on creating a time series analysis of COVID-19. This paper deep dives into using Python as a way to model the ARIMA model and capture the COVID-19 WHO's dataset.

Bibliography

[1] Vikas Kulshreshtha and N. K. Garg "Predicting the New Cases of Coronavirus [COVID-19] in India by Using Time Series Analysis as Machine Learning Model in Python". *The Institution of Engineers (India)* , 2020

[2] 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