HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

Analyzing the Effectiveness of COVID-19 Mitigation Policies Using ARIMA Forecasting

Ryan Hung[1] and Sandra Lee[#]

[1]Monte Vista High School, USA
[#]Advisor

ABSTRACT

The SARS-CoV-2 virus has triggered a worldwide pandemic situation which countries are desperately trying to adapt to. In order to halt the transmission of this virus, these countries have implemented COVID-19 mitigation policies, which are designed to suppress the spread and deadliness of the virus. However, there has not been much research into the effectiveness of these COVID-19 mitigation policies. Using data from the Kaggle Platform as well as the European Centre for Disease Prevention and Control, we hope to use an ARIMA time series forecasting model in order to identify effective COVID-19 mitigation polices. This will be done by analyzing the cases time series before the mitigation policy was implemented in a certain country and generating a predicted forecast curve during the time range of the mitigation policy. By comparing this generated curve with the actual curve, deviations will be able to be identified, indicating the significance of the mitigation policy during its implementation. Although most forecasting was relatively inaccurate due to a shortage in training data, one social distancing mitigation policy in South Korea had a clear deviation between the forecast curve without the influence of the mitigation policy and the actual curve. Overall, the ARIMA model has its merits and may prove to be useful with the collection of more data. By analyzing the effectiveness of these policies, future research into this topic may lead to a greater understanding about the transmission of COVID-19 and ways to suppress it.

Introduction

SARS-CoV-2 has left a devastating impact for most of the world due to its high transmission rates. This is in part due to the very nature of the virus, which is transmitted via the respiratory system. With a susceptible and connected human population, SARS-CoV-2 was able to quickly spread across the globe [1]. Various studies have been conducted in order to identify certain environmental factors that may aid in the transmission of SARS-CoV-2, such as differences in temperatures or economies [2]. In an attempt to control the spread of SARS-CoV-2, many countries implemented mitigation strategies that would decrease or eliminate the presence of these factors aiding in transmission. Although many factors have been identified, their use in these mitigation strategies has not been clearly documented. This paper will attempt to conduct research into using an autoregressive model in order to determine which COVID-19 mitigation strategies were effective. It first explains the use and pre-processing of each dataset, then discusses the model used, and finally explores the results found.

Literature Review

Transmission Factors

There have been numerous studies into candidate transmission factors of COVID-19.[2] Economic inequality, major sports events, and lower temperatures all contribute to the risk of COVID-19 transmission. Nutritional intake held a significant role as well, with the intake of vegetables, protein, Vitamin D, and Vitamin K all reducing COVID-19 risk, with increased alcohol intake increasing that risk. Other factors included age, sex, humidity, and urbanization level (Fig. 1).

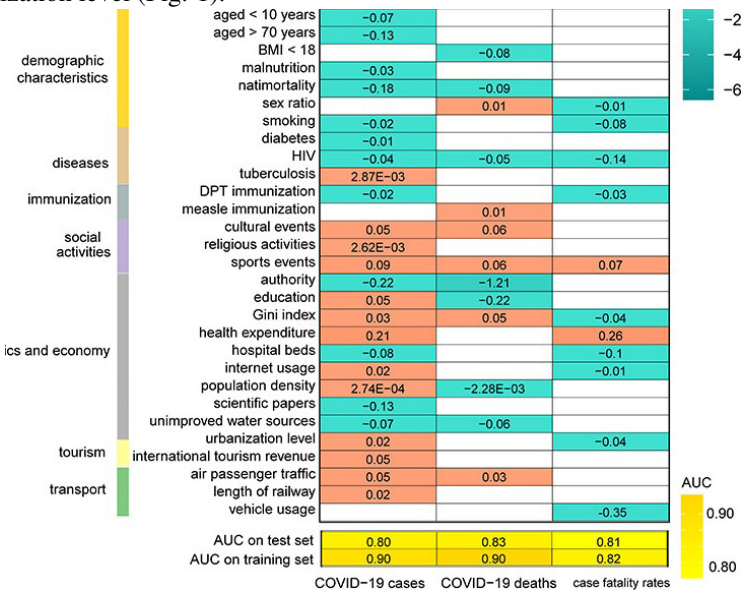

Figure 1. Novel Factors Associated with the Transmission/Fatality of COVID-19 [2]

Time Series Forecasting

Time series forecasting can be defined as utilizing a time series in order to predict future values for that series (Fig. 2). Being able to forecast data based off of previous time interval data has various uses, such as fundamental business planning. A more specific type of forecasting, Univariate Time Series Forecasting, makes predictions based solely on previous values of the time series [3].

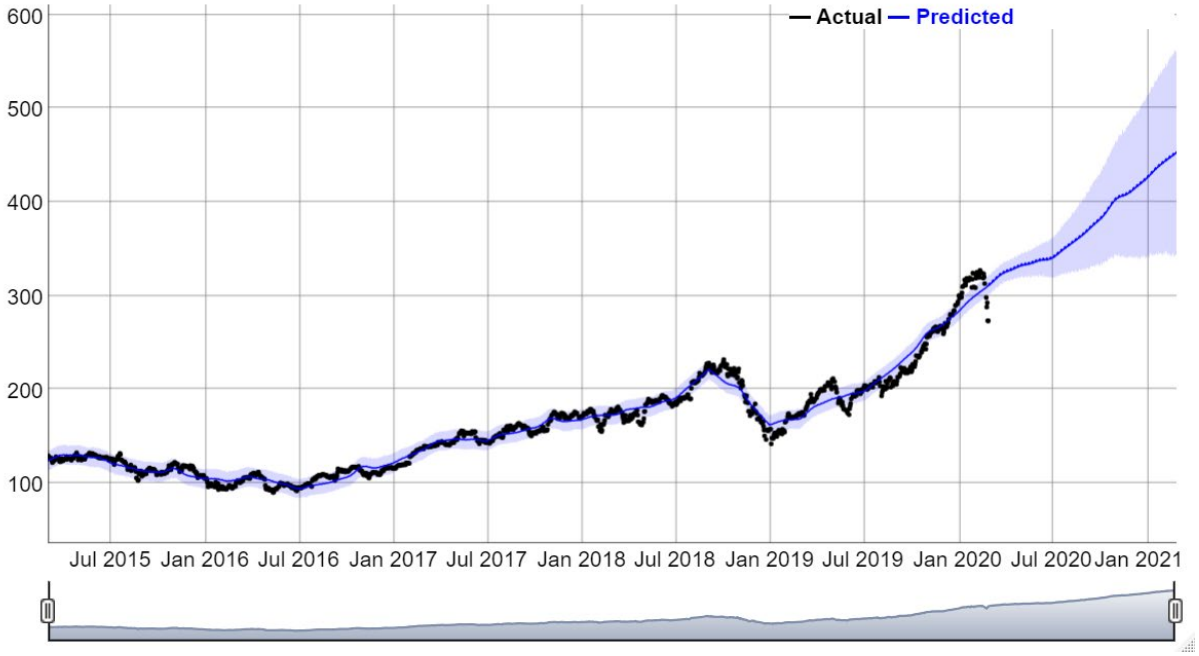

Figure 2. An Example of Time Series Forecasting [4]

ARIMA Model

ARIMA(Auto Regressive Integrated Moving Average) is a class of models that uses linear regression to attribute patterns or themes of a given time series based on its past values. This comes in the form of an equation that is created which can forecast future values. This model is characterized by 3 different terms or parameters. The ”p” or ”AR” term refers to the number of lag variables to be used as predictors. The ”q” or ”MA” term refers to the number of lagged forecast errors that should be used in the ARIMA Model. The ”d” term refers to the order of differencing needed in order to make the time series stationary. The model uses these three different parameters in order to explain the predicted values (Fig. 3) [3].

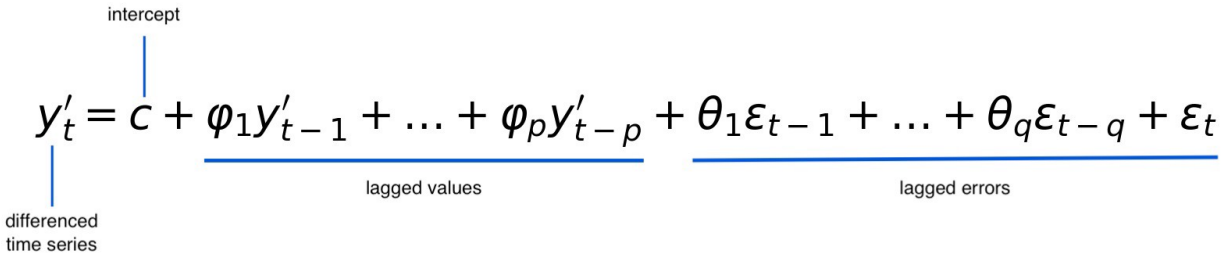

Figure 3. General ARIMA Model Equation [5]

Purpose

1. Determine if the auto-regressive model is effective when analyzing previously implemented COVID-19 mitigation strategies 2. Attempt to identify effective mitigation policies and their role in the reduced transmission 3. Rank the policies and analyze when they can be used by countries

Methodology

Data Acquisition/Pre-processing

The data used was obtained from the Kaggle platform (kaggle.com) as well as the European Centre for Disease Prevention and Control(https://www.ecdc.europa.eu/). The first dataset contained elements with the country name, the starting date, the intended ending date, a description of the COVID-19 mitigation policy in question, as well as keywords pertaining to the specific policy [6]. After pre-processing to remove non-existing values, the dataset contained 239 elements. The second dataset contained time series data with each country’s daily cases from December 31, 2019 to December 14, 2020 [7]. The data was stored in csv files, which were converted into Pandas Dataframes using Google Colaboratory.

ARIMA Model

The ARIMA model was trained on time series data originating from a single country at a time. These countries were determined by the availability of COVID-19 mitigation policy data for that country. Countries included China, South Korea, the United States, Denmark, Ireland, and Italy.

ARIMA Model ”AR” Parameter

In order to determine the ”AR” parameter, a partial auto-correlation plot was created in order to determine the correlation between each lag variable and the output variable. Any lag variables that crossed the significance limit were included in the model(Fig. 4) [3].

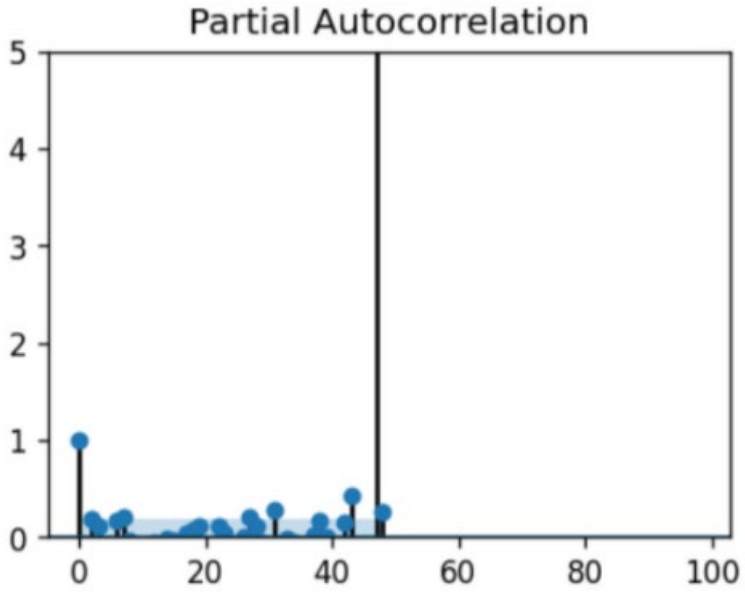

Figure 4. Example Partial Auto-correlation Plot

ARIMA Model ”MA” Parameter

In order to determine the ”MA” parameter, an auto-correlation plot was created in order to determine the correlation between each lag error term and the output variable. Any lag error(MA) terms that crossed the significance limit were included in the model(Fig. 5) [3].

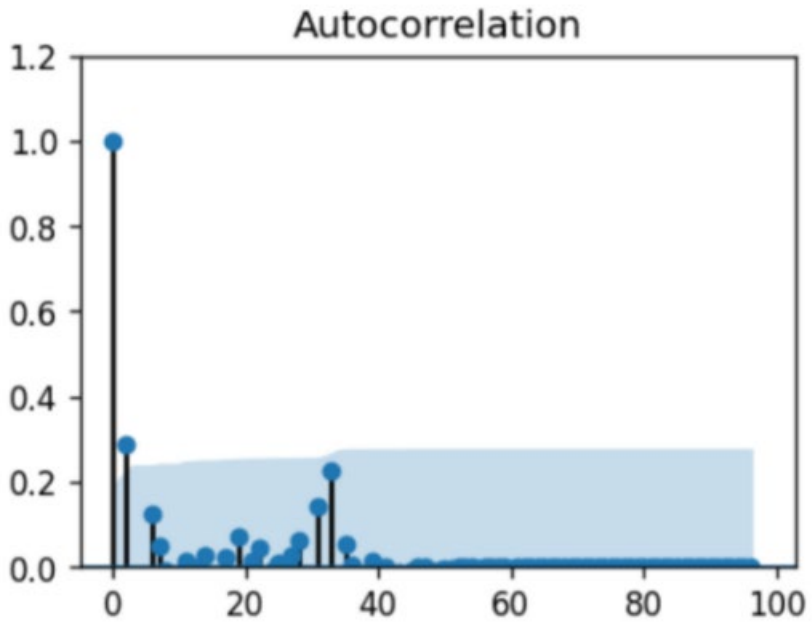

Figure 5. Example Auto-correlation Plot

ARIMA Model Differencing Parameter

In order to determine the order of differencing, the data was iteratively differenced until the time series became stationary. In all the cases, second order differencing was required in order for the series to become stationary, which was indicated by the auto-correlation plot roaming around a defined mean as well as approaching zero(Fig. 6) [3].

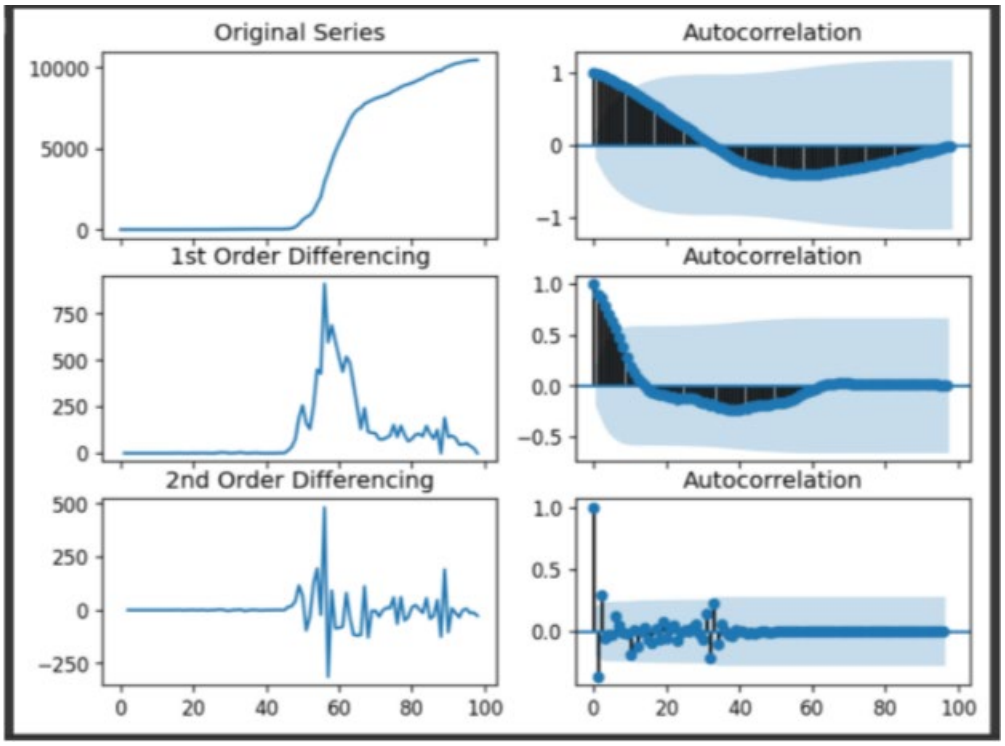

Figure 6. Example Differencing

Model Implementation

Once all parameters were obtained, the model was run using the ARIMA model class from statsmodels(https://www.statsmodels.org/stable/index.html). After being trained on the data, the model gave coefficients detailing how significant each term was in the model. If any parameters were deemed insignificant, the corresponding ”AR” or ”MA” value was changed.

	coef	std err	z	P>|z|	[0.025	0.975]
const	82.5376	82.068	1.006	0.317	-78.312	243.387
ar.L1.D.cases	0.3276	0.173	1.891	0.062	-0.012	0.667
ar.L2.D.cases	0.5711	0.156	3.663	0.000	0.266	0.877
ma.L1.D.cases	0.2957	0.195	1.513	0.134	-0.087	0.679

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0672	+0.0000j	1.0672	0.0000
AR.2	-1.6409	+0.0000j	1.6409	0.5000
MA.1	-3.3813	+0.0000j	3.3813	0.5000

Figure 7. Example Model Coefficients

Model Evaluation

The model was evaluated based on how well it fit the training data and whether there were any significant deviations. Residuals were plotted in order to determine constant mean and variance. The forecast values were examined for any significant deviations.

Results and Discussion

Results

For all of the model instances, the ARIMA model predicted values closely mirrored the training data. This was to be expected as the model itself was trained on that specific data.

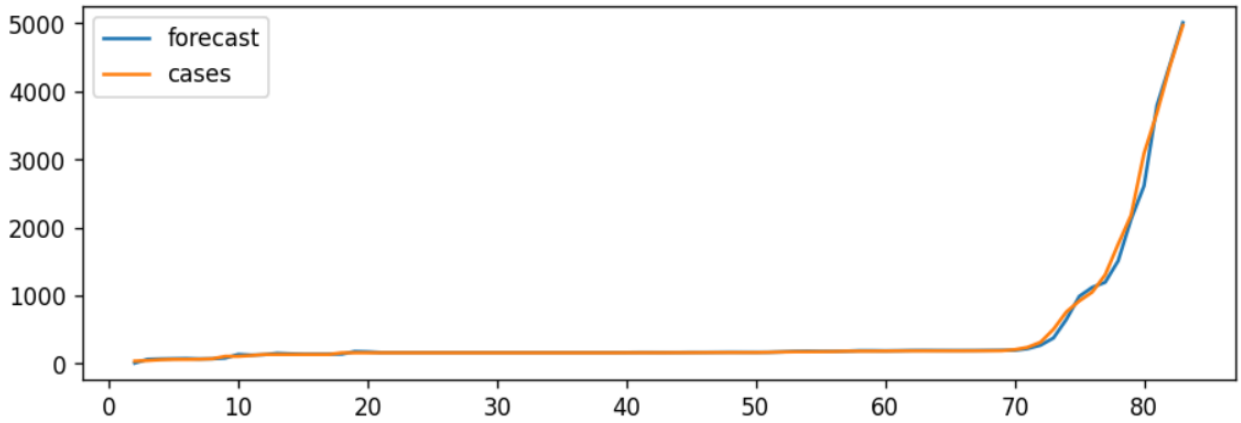

Figure 8. Model Predictions on Training Data

However, most forecasting data was unusable due to a general major spike happening that was usually not covered within the training data. This usually meant that the actual cases spiked past the 95 percent confidence interval band.

However, there was one instance in which the forecast was able to accurately predict within the confidence interval band. Interestingly, the model's forecast was able to predict a significant deviation due to South Korea's social distancing mitigation policy with a generated curve that was still within the confidence interval band.

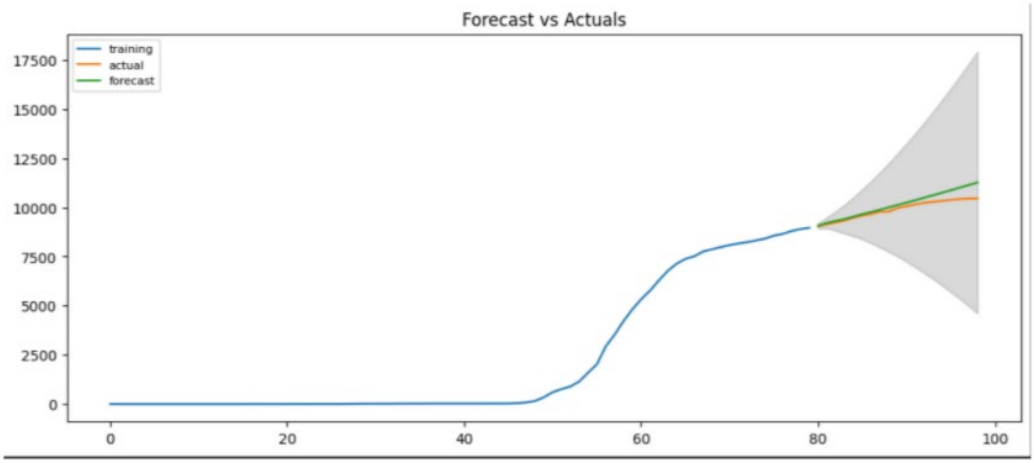

Figure 9. South Korea's Predicted Cases Curve Without Input from a Social Distancing Order

Analysis

Overall, the model itself fit the training data well, but failed to forecast specifically the March and April values due to the general spike in cases that happened to most countries during that time. Because most of the mitigation policies within the dataset were implemented within that time range, the training dataset generally did not include any indicator due to the spike. This unknown variable caused the model to generally forecast poor results. South Korea's generated curve was created without a spike variable to affect the testing data. Because the forecast curve followed the actual cases much more closely in that instance, it seems to imply that the model may be more successful given more data that accurately reflects the number of cases.

Conclusion

Summary

In conclusion, this model does seem to be able to accurately train itself on the cases time series data given enough data and proper parameters. However, the forecasting function results are inconclusive. Due to COVID-19 mitigation policies being implemented within March 2020, the model was forced to train on the months of January 2020 to March 2020. Within the time series, this data was not enough to predict the patterns past March 2020. In order for the model forecasting function to be properly tested, the model should be trained on COVID-19 mitigation policies implemented at a later time in order to expand the training dataset to include later, more influential months.

Future Investigations

Later COVID-19 mitigation policies

If more data on later COVID-19 mitigation policies is obtained, the model and its forecast function can be more extensively tested, and a more conclusive result can be achieved.

Deeper Analysis into Mitigation Factors

ong with data on the effectiveness of the policies, the analysis on the COVID-19 policies could also shed light on what factors these policies effect. Therefore, further research on this topic could possibly reveal hidden significant factors aiding in COVID-19 transmission.

Other Variants

The COVID-19 delta and omicron variant is a mutated version of the SARS-CoV-2 virus. Its new characteristics include a significantly more high transmission rate than the original including in individuals who have already had one vaccine dose[8]. Therefore, it may become imperative in the future that effective policies are implemented in order to suppress the transmission of the new delta variant. Further research could possibly reveal effective mitigation policies specifically tailored towards the suppression of the new variant.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Y. Yang, W. Shang, and X. Rao, “Facing the covid-19 outbreak: What should we know and what could we do?,” Journal of medical virology, 2020.
[2] M. Li, Z. Zhang, W. Cao, Y. Liu, B. Du, C. Chen, Q. Liu, M. N. Uddin, S. Jiang, C. Chen, et al., “Identifying novel factors associated with covid-19 transmission and fatality using the machine learning approach,” Science of
the Total Environment, vol. 764, p. 142810, 2021.
[3] S. Prabhakaran, “ARIMA Model – Complete Guide to Time Series Forecasting in Python,” 2019.
[4] J. LZP, “Time Series Forecasting With Prophet In R,” 2020.
[5] C. Bento, “Time Series Forecasting in Real Life: Budget forecasting with ARIMA,” 2020.
[6] P. Mooney, “COVID-19 containment and mitigation measures,” 2021.
[7] E. C. for Disease Prevention and Control, “Download historical data (to 14 December 2020) on the daily number of new reported COVID-19 cases and deaths worldwide,” 2021.
[8] K. Kupferschmidt and M. Wadman, “Delta variant triggers new phase in the pandemic,” 2021.