SARS-COV-2 Variant Analysis

Florida Polytechnic University

Ethan Bard

07 April 2022

Table of Contents

[**Abstract** 3](#_Toc100930568)

[Introduction 4](#_Toc100930569)

[Literature Review 4](#_Toc100930570)

[Data 10](#_Toc100930571)

[Methods 11](#_Toc100930572)

[Results 12](#_Toc100930573)

[Discussions 12](#_Toc100930574)

[Conclusions 12](#_Toc100930575)

[References 13](#_Toc100930576)

[Research links 17](#_Toc100930577)

# **Abstract**

* The abstract should be less than 450 words.
* The abstract is a summary of the thesis’s purpose, hypothesis, used methods, and results.
* The abstract section’s title (“Abstract”) should be title case, bold, underlined, and centered. Use a blank line to separate the title from the advisor(s) line.
* The advisors’ line contains the list of advisors and co-advisors separated by commas. Use a blank line to separate the advisors’ line from the body.

# Introduction

Since the emergence of SARS-COV-2, the novel coronavirus responsible for the COVID-19 pandemic beginning in early 2020, the World Health Organization (WHO) reports that nearly 500 million cases and over 6 million deaths have been documented (Who coronavirus (COVID-19) dashboard). As the virus has made its way across the globe, a multitude of mutations and variants have occurred, some of which resulting in significant changes in the contagiousness of the virus and the severity of the illness caused. It is expected that novel viruses go through many mutations in their early lifecycles, and as such it is essential that these variants are monitored to ensure the population is as prepared as possible (Katella, 2022). Furthermore, as countries around the world implement varying levels of disease prevention it is critical to be able to model and forecast cases, hospitalizations, and deaths in the near future to anticipate if additional actions need to be taken. In the following analysis, machine learning and time-series analysis methods will be demonstrated to evaluate their ability to classify, model, and forecast COVID-19 measures across the globe.

# Literature Review

**SARS-COV-2 Variants**

To better understand the challenges and difficulties in predicting cases of COVID-19 it is essential to discuss the effect of emerging novel mutations in the SARS-COV-2 virus. In an article by Tao et al. in 2021, a framework for understanding SARS-COV-2 variants is introduced by describing some of the most important features of SARS-COV-2 evolutions, and furthermore explaining the types of studies that will be required for the research, clinical, and public health communities to effectively manage the new threats introduced by emerging SARS-COV-2 variants (Tao et al., 2021). The article explains that emergence of SARS-COV-2 variants strongly affect the epidemiological and clinical aspects of the COVID-19 pandemic because variants can become more dangerous to the public by increasing rates of virus transmission, increasing the risk of reinfection, and reducing the protection available by currently available methods of treatment such as monoclonal antibodies and vaccinations. Additionally, these variants complicate the COVID-19 research agenda and increase the need for constant laboratory, epidemiological, and clinical research. So far, many of the variants that have been identified share specific mutations that all help to enable to virus to be able to spread and replicate even though many populations are increasing in immunity. According to Tao et al. this is due to the fact that many of the recently identified SARS-COV-2 mutations appear to antagonize the innate immune response to initial infection.

SARS-COV-2 variants are classified according to their lineage and component mutations (Tao et al., 2021). The two most commonly used naming conventions for SARS-COV-2 variants are the Phylogenetic Assignment of Named Global Outbreak (PANGO) lineage and the NextStrain systems, where the PANGO lineage is used more commonly due to its greater specificity. The format of the PANGO lineage system includes an alphabetical prefix and a suffix made from up to three numbers separated by periods which indicate the sub-lineages. An example of the PANGO lineage system is the notation for the Alpha variant which is “B.1.1.7”. Multiple variants can belong to the same lineage and are differentiated by their subsets of mutations. Primarily, variants are classified by their transmissibility, disease severity, and ability to evade humoral immunity. Tao et al. continues the article by describing some of the most prevalent variants identified at the time of the article’s writing. The ”Alpha” variant B.1.1.7, which was one of the first mutations of the SARS-COV-2 virus identified, accounted for the majority of infections in the United States and many European countries by the second quarter of 2021 (Tao et al., 2021). Compared to previous variants, many of which are not classified by PANGO, the Alpha variant was suggested to be 50% more transmissible and had an estimated 50% increased mortality. The “Beta” variant B.1.351 was first identified in South Africa between October 2020 and January 2021, where daily cases jumped from around 2,000 cases to 20,000 cases per day and subsequently spreading to the rest of the world. Again, compared to previous variants, this variant was estimated to be 50% more transmissible and be more likely to cause breakthrough cases in vaccinated individuals. The “Delta” variant B.1.617.2 originated in India along with “Kappa” B.1.617.1, both of which emerged from a common ancestor lineage. The Delta variant demonstrated unprecedented transmissibility, spreading to 54 countries and quickly became the most prevalent variant in the UK and USA. While SARS-COV-2 variants all differ in their transmission rates, disease severity, and risk of infection there is no significant evidence that suggest they respond differently to public health measures such as social distancing, personal protective equipment, or antiviral therapies. Therefore, Tao et al. claims that the most important aspect of emergent SARS-COV-2 variants is primarily their impact on vaccine efficacy.

A more recent article published on Yale Medicine by Kathy Katella, a senior clinical writer describes the features and implications of the newest and most prevalent SARS-COV-2 variant “Omicron” BA.1 and its subvariant BA.2. Katella explains that Omicron was first identified in Botswana and South Africa in late November 2021, and by December 2021 Omicron was the cause of United States cases skyrocketing to over a million (Katella, 2022). BA.2 has already completely overtaken BA.1 as the predominant variant in the United states due to its significantly more rapid transmissibility. Luckily, however, BA.2 does not appear to cause more severe disease than BA.1, and is also less severe than previous variants. Nonetheless, the CDC is currently classifying Omicron as a variant of concern in the US due to its extremely high transmissibility. At the time of this study’s writing, other variants that were being monitored including Alpha, Beta, Gamma, Epsilon, Eta, Iota, Kappa, Mu, and Zeta are no longer variants of concern because they are either no longer present in the US or are spreading slowly enough to not cause concern. Unfortunately, there is still a risk of further variants to continue to emerge due to limited access to vaccines around the world.

**Time-Series Analysis**

To aid the process of deciding how to design research and analysis, time-series data can be analyzed via cluster analysis to distinguish countries with different COVID-19 spread patterns and results. In a novel analysis performed by Zarikas et al. in 2020, countries were clustered with respect to active cases, active cases per population, and active cases per population and per area based on Johns Hopkins epidemiological data (Zarikas et al., 2020). In that analysis they claimed that clustering can support identification of possible causes of different impacts of the pandemic in different countries which in turn aids researchers to decide how to perform extended research. It was found that taking population and surface area into consideration resulted in significant changes to the cluster outputs. Some important conclusions made during their cluster analysis include the following: First, clustering with respect to active cases alone shows that countries in shared clusters have similar time evolution of the active cases, implying they have faced similar stresses to the health system. Second, clustering with respect to active cases per population shows that countries in shared clusters have experienced similar stresses to the society and the economy. Finally, clustering with respect to active cases per population per area is helpful for deriving conclusions about the impact of the disease that spreads more easily in densely populated areas. Furthermore, it was found that countries with the most critical situations tend to be smaller countries (Zarikas et al., 2020).

While cluster analysis is helpful to classify which countries have similar disease spread patterns, it is also important to be able to accurately forecast spread and results. The ability to accurately forecast when a surge of infections will hit its peak would significantly diminish the impact of the disease and allow officials to alter policies accordingly to plan ahead for preventative steps (Papastefanopoulos et al., 2020). An analysis conducted by Papastefanopoulos et al. in 2020 explored the performance of several popular time series modeling approaches for covid outbreak detection in ten countries that had the highest number of cases as of May 4th, 2020. Using data containing the progression of the virus and population of each country, six time series approaches were implemented and compared. The time series methods included in this analysis were ARIMA, HWAAS, TBAT, Prophet, DeepAR, and N-Beats. The countries included in the analysis were USA, UK, Italy, Spain, Russia, France, Turkey, Germany, Iran, and Brazil. To assess the performance of each time series model, an out of sample forecast was generated and the root mean squared error (RMSE) was calculated and compared against the other models. In summary, while each time-series approach has their own distinct positives and negatives, it was found that there was no “one size fits all” approach when it comes to predicting active cases for different countries; however, the ARIMA and TBAT approaches demonstrated superior performance in seven out of ten countries and achieved second best results in another two. Table 1 below presents the resulting RMSE of each time-series approach for each country. It is difficult to identify the specific reasons that certain algorithms perform better than others in one country but not in others, but the authors present a few suggestions that could provide some insight including country specific climate and geographical characteristics, population-related attributes such as population density, discrepancies in testing and measuring procedures and therefore data collection, and diversity in terms of quarantine and other social distancing measures (Papastefanopoulos et al., 2020).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **US** | **Spain** | **Italy** | **UK** | **France** | **Germany** | **Russia** | **Turkey** | **Brazil** | **Iran** |
| **ARIMA** | **0.007** | 0.080 | **0.006** | 0.005 | 0.061 | 0.006 | **0.002** | 0.004 | 0.004 | 0.003 |
| Prophet | 0.014 | 0.065 | 0.019 | 0.008 | 0.044 | 0.037 | 0.015 | 0.045 | 0.009 | 0.016 |
| HWAAS | 0.173 | 0.031 | 0.007 | 0.004 | 0.011 | 0.005 | 0.002 | **0.001** | 0.006 | 0.001 |
| NBEATS | 0.037 | 0.050 | 0.009 | 0.038 | **0.004** | 0.013 | 0.027 | 0.018 | 0.011 | 0.004 |
| Gluonts | 0.045 | 0.109 | 0.044 | 0.046 | 0.011 | 0.058 | 0.034 | 0.094 | **0.003** | 0.002 |
| **TBAT** | 0.010 | **0.029** | 0.006 | **0.004** | 0.007 | **0.003** | 0.002 | 0.002 | 0.006 | **0.000** |

Table 1: RMSE Results of Time-Series Model Approaches by Country

The Auto Regressive Integrated Moving Average (ARIMA) model was originally designed for economic application and provides a high level of interpretability because the relationship between the independent variables and dependent variables are easily understood and simple to explain. The ARIMA model assumes a linear correlation between the time-series values and attempts to exploit these linear dependencies in observations to extract patterns, while removing high frequency noise from the data (Papastefanopoulos et al., 2020). A significant advantage to the ARIMA method is that models can be performed in an automated way to maximize prediction accuracy. For these reasons, the ARIMA method will be the time-series method of choice for the analysis conducted in this report.

# Data

The data collected for this analysis consists of two time-series datasets representing international measurements of several features related to COVID-19 and are compiled in weekly and daily intervals. The primary dataset includes daily measurements of features such as number of cases, hospitalizations, deaths, and vaccinations. Each of the listed measurements are provided in multiple formats including raw daily counts, cumulative totals, smoothed daily/weekly counts, and smoothed daily/weekly counts per hundred thousand or per million. The data includes records ranging from as early as January 1st, 2020 and is regularly updated with current observations. These records are compiled and provided by the team at Our World in Data. Our World in Data (OWID) is an organization of researchers, data scientists, and engineers whose goal is to “publish the research and data to make progress against the world’s largest problems” (Roser). OWID primarily brings data together from four types of sources including specialized institutes, research articles, international institutions or statistical agencies, and official data from government sources (Roser).

The secondary dataset utilized in this analysis includes weekly international measurements of COVID-19 sequencing results. The features provided in this dataset include the total number of sequences analyzed, total number of sequences classified per variant, and the proportion of sequences classified per variant as a percentage of the total number of sequences. The data does not directly represent the number of COVID-19 cases but provides insight as to which COVID-19 variant(s) are the most prevalent internationally at a given point in time. These records are compiled and provided by GISAID ranging from December 29th, 2019, to the current day. GISAID is a global science initiative and primary source established in 2008 that provides open access to genomic data of influenza viruses and the coronavirus responsible for COVID-19. This includes “genetic sequence and related clinical and epidemiological data associated with human viruses, as well as species-specific data associated with avian and other animal viruses, to help researchers understand how viruses evolve and spread during epidemics and pandemics” (Mission).

# Methods

* This chapter will provide details on the chosen methods, designs, measures, and philosophy behind these choices. In addition, this chapter should include a description of any conduct experiment.
* (Chyon et al., 2021) Time series analysis and predicting COVID-19 affected patients bsay ARIMA model using machine learning

# Results

* This chapter contains the result of the thesis. If possible, organize the thesis’s results into figures. Otherwise, organize the results into tables. Finally, divide the results into sections and subsections based on the research questions they address.

# Discussions

* This chapter contains the analysis, explanations, and discussions of the results. It should also include statements whether the results support the hypothesis or not, with some reasoning if it does not.

# Conclusions

* This chapter should be a summary of the study indicating whether the study met its goals or not.

# References

Amidon, A. (2021, August 27). *How to apply K-means clustering to time series data*. Medium. Retrieved April 13, 2022, from https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3

Brownlee, J. (2020, August 14). *A gentle introduction to autocorrelation and partial autocorrelation*. Machine Learning Mastery. Retrieved April 13, 2022, from https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/

Centers for Disease Control and Prevention. (n.d.). About covid-19. Centers for Disease Control and Prevention. Retrieved April 13, 2022, from https://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19.html

Centers for Disease Control and Prevention. (n.d.). CDC's role in tracking variants. Centers for Disease Control and Prevention. Retrieved April 13, 2022, from https://www.cdc.gov/coronavirus/2019-ncov/variants/cdc-role-surveillance.html

Centers for Disease Control and Prevention. (n.d.). What is genomic surveillance? Centers for Disease Control and Prevention. Retrieved April 13, 2022, from https://www.cdc.gov/coronavirus/2019-ncov/variants/genomic-surveillance.html

Chyon, F. A., Suman, M. N. H., Fahim, M. R. I., & Ahmmed, M. S. (2021, December 14). *Time series analysis and predicting COVID-19 affected patients by Arima model using machine learning*. Journal of Virological Methods. Retrieved April 13, 2022, from https://www.sciencedirect.com/science/article/pii/S0166093421003724

Frost, J., mustafa, G., & Buenaobra, B. (2021, May 18). *Autocorrelation and partial autocorrelation in time series data*. Statistics By Jim. Retrieved April 13, 2022, from https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/

Halford, E. A. (2021, March 10). *Setting Arima model parameters in R: Grid search vs. auto.arima()*. Medium. Retrieved April 13, 2022, from https://towardsdatascience.com/setting-arima-model-parameters-in-r-grid-search-vs-auto-arima-19055aacafdf

*History*. GISAID. (n.d.). Retrieved April 13, 2022, from https://www.gisaid.org/about-us/history/

*How to perform Correlation Analysis in time series data using R? - luba*. LOB.DATA. (n.d.). Retrieved April 13, 2022, from https://www.lobdata.com.br/2020/09/15/how-to-perform-correlation-analysis-in-time-series-data-using-r/

Katella, K. (2022, March 30). Omicron, Delta, Alpha, and more: What to know about the coronavirus variants. Yale Medicine. Retrieved April 13, 2022, from https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron#:~:text=Delta%20(B.,its%20place%20in%20mid%2DDecember.

Mission. GISAID. (n.d.). Retrieved April 13, 2022, from <https://www.gisaid.org/about-us/mission/>

Papastefanopoulos, V., Linardatos, P., & Kotsiantis, S. (2020, June 3). Covid-19: A comparison of time series methods to forecast percentage of active cases per population. MDPI. Retrieved April 14, 2022, from https://www.mdpi.com/2076-3417/10/11/3880/htm

Roser, M. (n.d.). About. Our World in Data. Retrieved April 13, 2022, from https://ourworldindata.org/about

Tao, K., Tzou, P. L., Nouhin, J., Gupta, R. K., de Oliveira, T., Kosakovsky Pond, S. L., Fera, D., & Shafer, R. W. (2021, September 17). The biological and clinical significance of emerging SARS-COV-2 variants. Nature News. Retrieved April 14, 2022, from https://www.nature.com/articles/s41576-021-00408-x

World Health Organization. (n.d.). Who coronavirus (COVID-19) dashboard. World Health Organization. Retrieved April 13, 2022, from https://covid19.who.int/

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., & Zervas, E. (2020, May 29). Clustering analysis of countries using the COVID-19 cases dataset. Data in Brief. Retrieved April 14, 2022, from https://www.sciencedirect.com/science/article/pii/S2352340920306818

# Research links

<https://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19.html>

<https://www.lobdata.com.br/2020/09/15/how-to-perform-correlation-analysis-in-time-series-data-using-r/>

<https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/>

<https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/>

* Really good explanations about how to interpret the correlation plots

<https://towardsdatascience.com/setting-arima-model-parameters-in-r-grid-search-vs-auto-arima-19055aacafdf>

* R arima grid search method

<https://www.sciencedirect.com/science/article/pii/S0166093421003724>

* COVID forecasting with arima

<https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3>

* K-means Clustering time series

<https://cran.r-project.org/web/packages/Rssa/Rssa.pdf>

* R SSA documentation

<https://www.researchgate.net/publication/228092069_Basic_Singular_Spectrum_Analysis_and_Forecasting_with_R>

* R SSA tutorial and explanation

<https://www.cdc.gov/coronavirus/2019-ncov/variants/genomic-surveillance.html>

* Cdc explanation on variants

https://www.nature.com/articles/s41576-021-00408-x