SARS-COV-2 Variant Analysis

Florida Polytechnic University

Ethan Bard

07 April 2022

Table of Contents

[Abstract 3](#_Toc100747666)

[Introduction 4](#_Toc100747667)

[Literature Review 4](#_Toc100747668)

[Data 4](#_Toc100747669)

[Methods 5](#_Toc100747670)

[Results 5](#_Toc100747671)

[Conclusions 5](#_Toc100747672)

[References 6](#_Toc100747673)

[Research links 7](#_Toc100747674)

# **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.
* Analysis of worldwide covid data examining variant effect on various trends
* Major variables of interest:
  + Cases
  + Hospitalizations
  + Deaths
  + Vaccinations
  + Variant (% of sequences)
* Time Series Analysis
  + Clustering
  + Modeling
  + Forecasting
    - ARIMA
    - Multivariate Regression
* Goal: Identify countries with emerging variants and forecast cases, deaths, and variant (%)
* Goal: Forecast cases and identify countries still seeing increased cases
* Goal: Cluster trends and examine those that stand out

# Introduction

* The introduction chapter should emphasize the purpose of the study and summarize the background and importance of this research. This chapter should also clearly state the hypothesis and the objectives of the research. Finally, this chapter should introduce the user to the outline of the thesis.

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

* The literature review will summarize the existing research in the field with references to these research studies and their authors. At the end of the literature review, clearly state the identified gaps in the existing research solutions and address these gaps by this thesis. In other words, you are introducing the reader to your work in the next chapter.

# 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.

# 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/

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

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

# 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