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.

**Variant Analysis**

(Tao et al., 2021) The biological and clinical significance of emerging SARS-CoV-2 variants

* The emergence of sars cov 2 variants with novel spike protein mutations influence the epidemiological and clinical aspects of the covid-19 pandemic.
* Variants can increase rates of virus transmission and/or increase the risk of reinfection and reduce the protection afforded by neutralizing monoclonal antibodies and vaccination
* Variants complicate the covid-19 research agenda and necessitate additional avenues of laboratory, epidemiological, and clinical research.
* The increasing number of sars cov 2 variants share a repertoire of mutations that is enabling the virus to spread despite rising population immunity while maintaining or increasing its replication fitness.
* Whereas most emerging mutations reduce the protective effects of neutralizing antibodies generated by infection and vaccination, several recently identified mutations appear to antagonize the innate immune response to initial infection
* Initially, genetic sequencing suggested that sars cov 2 was exceptionally well adapted to humans, spreading rapidly with little evidence for natural selection among circulating viruses. This changed during the later months of 2020 with the first reports of emergent sars cov 2 variants associated with increased transmissibility, disease severity, and escape from humoral immunity (immunity mediated via host antibodies including those that directly neutralize virus as well as those that recruit other host immune functions)
* **In this review we create a framework for understanding sars cov 2 variants by describing fundamental aspects of sars cov 2 evolution, then describe the biological properties and epidemiological characteristics, then describe the types of study required for the research, clinical and public health communities to respond to the new threat posed by emerging sars cov 2 variants.**
* The phylogenetic classification of emergent sars cov 2 lineages have been difficult because new lineages often differ from one another by just a few nucleotides.
* Geographical classification has been challenging because most variants have been detected in multiple countries and there are marked disparities in the proportion of viruses undergoing sequencing in different countries.
* Two commonly used systems have been developed for epidemiological surveillance: the Phylogenetic Assignment of Named Global Outbreak (PANGO) lineage and the NextStrain systems. The PANGO lineage system provides greater specificity and is used more frequently. It contains an alphabetical prefix and a suffix containing up to three numbers separated by periods indicating sub-lineages (ex: B.1.1.7)
* Currently circulating sars cov 2 VOCs and VOIs share several mutations that enable them to spread in the face of rising population immunity while maintaining or increasing their replication fitness.
* Variants are classified according to their lineage and component mutations.
* Viruses belonging to the same lineage but containing different subsets of mutations can be classified as different variants. Variants are classified by their transmissibility, disease severity, and ability to evade humoral immunity.
* Increased transmissibility is demonstrated by the ability of a variant to outcompete other variants and to display a higher effective reproduction rate and/or secondary attack rate compred to other circulating variants.
* Disease severity has been assessed using mortality data and rates of hospitalization.
* Variants associated with higher virus levels may be more transmissible and/or cause more severe disease.
* Alpha (B.1.1.7) accounted for the majority of infections in the USA and many European countries by the second quarter of 2021, studies suggest that it was approximately 50% more transmissible than previously circulating UK variants. It was also associated with threefold to eightfold higher upper-airway levels and an estimated 50% increased mortality.
* Beta (B.1.351) Between October 2020 and January 2021, daily cases in south Africa increased from approximately 2,000 to more than 20,000 reported cases per day. This increase occurred in a setting in which more than 30% of the population was estimated to have already been infected and was associated with the emergence of the Beta variant. This variant was estimated to be 50% more transmissible than the linneages that preceded it. This variant has been associated with reduced vaccine efficacy.
* Delta (B.1.617.2) Originating from India in early 2021, Delta and Kappa(B.1.617.1) diverged from a common ancestor. The Delta variant demonstrated increased transmissibility, spreading to 54 countries and rapidly replacing the alpha variant in the UK and USA.
* Although sars cov 2 variants differ in their transmission rates, disease severity and risk of reinfection, there is noe evidence that they are differentially affected by non-pharmaceutical public health measures such as social distancing and the use of personal protective equipment, or that they will respond differently to most antiviral therapies.
* **The most important consequence of emergent sars cov 2 variants, therefore, is their impact on vaccine efficacy.**
* As the spectrum of variants is expanding and shifting faster than epidemiological studies can be conducted, laboratory correlates of protection against sars cov 2 variants have become a high priority.
* **Therefore, a strategy that combines genomic surveillance, in vitro neutralization studies, and vaccine efficacy studies should be maintained to identify those variants that pose the greatest threat to current vaccines and to guide the development of immunogens for second-generation vaccines.**

(Katella, 2022) Omicron, Delta, Alpha, and more: What to know about the coronavirus variants

* Omicron (BA.1) and its subvariant BA.2 which surpassed the original omicron strain in March 2022 to become the predominant variant in the U.S. Omicron was first identified in Botswana and South Africa in late November 2021, and cases quickly began to surface and multiply in other countries. By December, Omicron was causing daily case numbers in the U.S. to skyrocket to over a million.
  + The BA.2 subvariant appears to spread more easily than BA.1 but does not appear to cause more severe disease than BA.1
* Omicron is more transmissible than delta was, but is less severe than previous variants.
* The CDC is currently focused on Omicron as a variant of concern in the US, which is a classification given to variants that show increased transmissibility, could cause more severe disease, may be resistant to antibodies from previous infections or vaccinations, and/or show an ability to evade diagnostic detection.
* Other variants being monitored such as Alpha, Beta, Delta, Gamma, Epsilon, Eta, Iota, Kappa, Mu, and Zeta either are no longer detected in the US or are spreading at a slow enough pace that they don’t pose a serious risk.
* Experts say one of their concerns is that limited access to vaccines around the world will drive surges in COVID-19 cases, and this will increase the chances that concerning variants will continue to emerge.

**Time Series Analysis**

(Zarikas et al., 2020) Clustering analysis of countries using the COVID-19 cases dataset

* Novel analysis which results to clustering countries with respect to active cases, active cases per population, and active cases per population and per area based on Johns Hopkins epidemiological data.
* The presented cluster results could be useful to a variety of different policy makers, and suggests a new specially designed clustering algorithm adapted to the request for comparison of the various covid time-series of different countries.
* This data is useful because various countries can be clustered to objectively distinguish countries with different COVID-19 spread and results
* Clustering can be further expected to support the identification of possible causes of these different impacts of the pandemic in different countries, thus helping researchers to decide how to design more extended research.
* Taking population into consideration significantly affected the clustering results
* Surface area of countries also affected the cluster results
* Countries with the most critical situations tend to be small countries.
* Clustering with respect to active cases means that the elements of these clusters are countries that have similar time evolution of the active cases, which means they have faced similar stresses to the health system
* Clustering with respect to active cases per population means that the countries that belong to the same cluster have experienced similar stresses to the society and the economy.
* Clustering with respect to active cases per population per area is useful for driving conclusion about the impact of the disease that spreads more easily in densely populated areas (countries with dense cities are more vulnerable)

(Papastefanopoulos et al., 2020) COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population

(Chyon et al., 2021) Time series analysis and predicting COVID-19 affected patients bsay ARIMA model using machine learning

# 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/>

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