Coronavirus Disease 2019 Data Analysis

Junhe Zhang, M.S.

Nima Zahadat, Ph.D.

The George Washington University

**ABSTRACT**

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. By April 2021, more than 149 million people get infected by COVID-19 and caused more than 3.14 million fatal cases, known as the most dangerous infectious diseases in the last 20 years. The purpose of this project was to demonstrate data mining and machine learning techniques using the record data of COVID-19 cases in the worldwide range for the spread pattern analysis of the pandemic and forecast the future new positive case trends in the United States of America. Modeling focused on the following data features: report date, number of populations, latitude and longitude, the average age of the people, GDP of the countries or states, total area space, and land area space, and COVID-19 cases (positive cases and death cases). The best model was the Random Forest Regressor which was selected using the Hyperparameter Tuning technique to correctly predict the people positive new cases 94% of the time. The feature with the most information gained as assessed by Random Forest Regression was GDP and report date. Additionally, Linear Regression was used to interpret each feature’s positive or negative effects on the model with 35% accuracy. Finally, use the advanced time-series analysis GARCH model to predict the future positive new cases trend for each state in the United States of America to assist medical resources arrangement and as a reference to the future trend of the spread of infectious disease.

Keywords: COVID-19, machine learning, Hyperparameter Tunning, Random Forest Regression, Linear Regression, Time-Series Analysis.

**INTRODUCTION**

The first cases of a new contagious disease were diagnosed in early December 2019 in the city Wuhan, the capital of Hubei province of the People’s Republic of China, and spread worldwide at a breakneck speed. Within a short period, the outbreak developed into a severe pandemic around the world that infects millions of people. Eventually, the novel disease was named coronavirus disease 2019 (COVID-19), and the new virus was identified as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). By April 2021, COVID-19 had caused more than 149 million positive cases and more than 3.14 million fatal cases worldwide, known as the most severe epidemic since economic globalization (Pitlik & Silvio Daniel, 2020, table 4).

Economic globalization has been developing faster and faster in recent decades, which has become an irreversible trend. This trend has blurred the boundaries between nations and brought people from different geographic areas closer together. This global market pattern has created enormous benefits and will be intensified more and more in the future. However, this pattern has also made serious potential health risks worldwide and put a new and inevitable challenge to human epidemic prevention, which is how to control the epidemic on a global scale. The global spread of the epidemic will not cease until the outbreak in all countries participating in the global market is under control.

As the second-largest economy globally, China has a considerable share in the international market, so the COVID-19 epidemic with China as its source would have such a significant impact across the globe. The COVID-19 spread over 200 countries across five continents in just a week and has become a substantial threat to human health. This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment. In contrast, many healthcare workers themselves have been infected (Yazeed, Shira & Noam, 2021). Thus, the capacity to correctly forecast the future trend of positive new cases caused by a pandemic is critical and would significantly affect helping local governments determine healthcare resource arrangements and help to prevent the epidemic from rising in advance. The data collected from this pandemic is precious and must be studied in all aspects to provide help for humanity to fight against future epidemics.

**LITERATURE REVIEW**

The SARS-CoV-2 virus is a highly contagious novel virus. Viruses account for only a tiny fraction of the 1400 or more different pathogen species that plague humans. Examples are bacteria, fungi, or helminths. However, recent AIDS and influenza pandemics illustrate that viruses should be high on the human global public health threat list. On the other hand, over the past decades, more and more newly recognized human pathogens were caused by viruses. In Pitlik and Silvio’s study of “COVID-19 Compared to Other Pandemic Diseases”, five out of Six of the most recent emerging infectious disease pandemics have involved viruses, such as AIDS (RNA 1981-present), SARS (RNA 2002-2003), Ebola (RNA 2014-2016), MERS (RNA 2015-present), and the current COVID-19 (RNA 2019-present).

In the study of Mark, Kyle and Liam of RNA viruses, viruses can be categorized into two major types: RNA viruses and DNA viruses. In Mark, Kyle, and Liam’s study of RNA viruses, the latter DNA viruses primarily consist, except for a handful of pox and herpes viruses, or viruses that have probably been present in and coevolved with humans for long periods. RNA viruses, on the other hand, are mostly zoonotic, meaning that they can infect both human and vertebrate animals. The article also states that many RNA viruses, not regarded as zoonotic, are believed to have had recent (in evolutionary terms) zoonotic origins. Even if the virus is contained in human society, it can spread and evolve in other vertebrates and re-spread to humans. This characteristic of RNA viruses makes them one of the most challenging viruses to eradicate and prevent.

The current global pandemic caused by SARS-CoV-2 was a typical example of zoonotic RNA virus transmission. In the report published by the World Health Organization (WHO), the virus was originated in bats, and human transmission primarily occurs through direct, indirect, or close contact with infected people through infected secretions such as respiratory secretions, saliva or through respiratory droplets that are expelled when an infected person coughs, sneezes, or speaks. Therefore, it is vital to be able to document the spread of the epidemic quickly and accurately. At the national level, many governments established a digital surveillance system to keep track of international travelers, COVID-19 positive contacted people, and those with quarantine status. Besides this, many other independent agencies also keep a close watch on COVID-19 development, such as John Hopkins University, Centre for Disease Control, NIH, etc.

“The effective surveillance from local community level to the regional level to the national level to world level is the key to find effective solutions for containment of the virus. Worldwide surveillance also provides on-time alerts as to potential locations where the infection may be emerging” (Srivastava N et al. 2020). This statement confirms that the effective use of monitored epidemic data to predict the trend of the future pandemic can help to accelerate the process of containing the virus. Moreover, studying the spread of the COVID-19 epidemic to produce valuable insights or machine learning models helps predict the spread pattern of the future contagious virus.

**RESEARCH METHODOLOGY**

Python 3.9 and standard packages (pandas, NumPy, seaborn, plotly) were used to process the raw .csv file and analyze the data. Data preprocessing techniques were used to merge multiple online scraped data into the primary dataset and then group the samples by area region and report date sample data. The combined dataset is standardized for boosting the training process for selected machine learning models. Exploratory Data Analysis (EDA) provided valuable insights about the dataset, help to discover the underlying patterns through data visualization of variable distributions and variable trend patterns across different regions. Later, the hyperparameter tunning technique is applied to evaluate the best model among the candidate models by comparing their R-squared score and select the model with the highest score. The candidate models are Random Forest Regressor, ElasticNet Regressor, XGBoost Regressor, and SGD Regressor. The Random Forest Regressor is used to discover the level of importance of each independent variable in quantification, then use other models to determine the positive or negative effects each independent variable contributed to the target variable. Finally, use the GARCH model, a time-series analysis algorithm, to predict the possible future trend of the target variable for different regions. In the end, analysis and conclusions were drawn based on the results of all processes in the project.

**DATA**

The COVID-19 Activity data is a dataset of real-time monitoring of COVID-19 epidemic data worldwide. It keeps tracking all the activating COVID-19 cases, including new positive and death cases and much more. This dataset is an aggregated dataset that includes any data that epidemic surveillant systems have monitored from nearly all possible sources, for example, World Health Organization, European Centre for Disease Prevention and Control, US CDC, WorldoMeters, and many more. This precious dataset is started and maintained by the dedicated, hard-working scholars in Johns Hopkins University Center for System Science and Engineering, also supported by ESRI living Atlas Team. The initial COVID-19 Activity dataset contained 13 factors and 1,048,576 samples, including the daily COVID-19 new cases data for countries affected by the outbreak. The epidemic data for some countries are detailed to the counties such as the United States of America and the People’s Republic of China. Supplementary datasets like countries’ geographic data and demographic data are merged into the COVID-19 Activity dataset. The datasets are then split into two groups which are country-level group and state-level group so that the researchers could observe the problem from a macro and micro perspective. The final dataset after Data Preprocessing and EDA phases contained 18 factors and 50,809 samples. Some further data processing was done along with the model fitting stage.

**DATA ANALYSIS**

The columns for Country’s Alpha 2 and 3 code and the Data Source were dropped immediately from the dataset. Some countries with insufficient (less than 100) data samples were dropped as well. Any Null sample points cause by datasets merging were handled using the forward filling technique, which used the most recent data point to fill the missing data for the current date. Columns like Country’s Name and Continent’ name were used only during the data visualization phase. The coordination data like the Latitude and Longitude of the corresponding geographic region would be used as substitutions to train the machine learning model. Visualization, including treemap, help to demonstrate the portion of COVID-19 cases of each country in the world, heatmap, used to display the intensity of COVID-19 new cases in through time for each country. The visualizations were generated with matplotlib, seaborn Python packages, and Tableau application. The Data Modeling phase was done using both the sklearn package and the Garch package. The sklearn package support splitting and scaling the data, fitting the training data, and predicting the outcomes of test data. The best two models that were determined using the Hyperparameter Tuning technique were Random Forest Regressor and ElasticNet Regressor; they were used to score all the features’ importance within the model’s attributes. From the perspective of time series analysis, the Garch model is used to predict the trend of COVID-19 new cases in the next 90 days in each state in the United States starting on April 2.

**KEY FINDINGS**

Through Exploratory Data Analysis (EDA) phase, several informative plots were generated to better understand the features and patterns of the data being utilized. On a global scale, treemap showed an intuitive demonstration of the portion of accumulative COVID-19 positive cases for each country by April 2. The continent most affected by the pandemic in the Americas, followed closely by Europe and Asia. The United States had the highest number of infections globally, with a total of 30,631,700 positive cases. The countries in second and third place were Brazil and India, which had 12,910,082 and 12,392,260, respectively. Except for Europe, the three most severely infected countries on other continents accounted for more than one-third of the total number of infections on the continent. Europe showed a different pattern. Nine of the ten most affected countries are members of the European Union, and there was no significant difference in the number of infections in each country. This phenomenon may be caused by the frequent trade exchanges between each E.U. member state; On the other hand, all these severely infected countries had the highest GDP in each state. All evidence indicated that frequent trading activity during the pandemic is hazardous. Chart, treemap chart

Description automatically generated

*Figure 1*. Treemap for each country on a global scale.

Within the continental U.S, the pandemic began to break out in large states with dense populations and high GDP like New York and California in mid-March and then quickly spread to the rest of the inland states in two weeks. In terms of the number of infections, these states are also far ahead of other states; this result corroborated the ideas obtained from global scaled data, which was, GDP and population had a positive effect on the number of infections. However, from the point of view of infection rate, the result seemed to violate the conclusions obtained from the previous section. The top 10 states with the most infection rate (the number of infections divided by the total number of people) were deficient in GDP and population rankings. After careful observation, these states had one thing in common: they are all states that voted for the Republican Party in the 2016 election. In general, the average infection rate of the states that voted for the Republican Party (red states) and Democrat Party (blue states) was 10.24% and 7.75%, respectively, which was 2.5% higher.

Chart

Description automatically generated

*Figure 2*. Rank of infection rate statewide within USA

Next, two sets of boxplots were created to demonstrate the spread of the pandemic globally and within the United States from a time-series perspective. On a global scale, the outbreak in Asia started in January and was brought under control after four weeks of continuous growth. After a quiet period of three weeks, the epidemic in Europe began to increase rapidly in mid-March; the outlier of boxplots indicated the spread of the epidemic in Europe far exceeds that of other continents. Like Asia, the pandemic in Europe was effectively controlled after four weeks. Beginning in the third week of the epidemic in Europe, the epidemic in the Americas began to emerge on March 22; the spread of the epidemic was slowed after a three-week growth period but did not get controlled until mid-July. There were four peaks in the spread of epidemics on all continents: the four-week period of the first outbreak of the epidemic, the outbreak period in autumn, the outbreak period during Thanksgiving, and the outbreak period in spring. The prime time to contain the epidemic was in the first four weeks; examples were China, Japan, South Korea, European Countries. These countries had successfully contained the spread of the epidemic during the prime time. In nearly six months, the spread of the epidemic has been significantly reduced until the second peak in autumn began. Figure 3 has been provided as an example.

Chart, histogram

Description automatically generated

*Figure 3*. Epidemic Trend global scale

The spread pattern of the epidemic in the United States closely related the four peaks pattern that was discussed earlier. In addition, the blue state was affected by the epidemic about a week earlier than the red state, and then the government started to execute the Stay-at-Home order in mid-March. This action has significantly slowed the spread of the epidemic. In most states, the spread of the epidemic was controlled within the next six months. However, the red states were slowly losing control of the epidemic, with many of them have risen to between the 3rd and 4th Inner Quartile Range (IQR). This situation became much evident after the third peak, with many of the red states becoming outliers. Figure 4 has been provided as an example.

Chart

Description automatically generated

*Figure 4*. Epidemic Trend United States scale

The two models chosen for this dataset were Random Forest Regressor and ElasticNet Regressor. Random Forest Regressor was the model with the highest R-squared score that was selected using Hyperparameter Tunning with 94.8% accuracy. ElasticNet Regressor was selected to explain the effect that each independent variable on the target from a linear perspective with 30.6% accuracy. The coding process is similar for both. The dataset must be split into training and testing dataset, scaling the data could significantly boost the time of the training process, remove or encode all the categorical variables, setup the hyperparameters for all candidate models, and fit the model using Grid-Search Cross-Validation to get the model with the best combination of parameters. Although Random Forest Regressor had 94.8% accuracy, it was not a good predictor for future epidemic trends due to its characteristic of the classifier. It was used to analyze the influence of each variable in the model.

Through the integration of the two model coefficients, the importance of the feature was shown as a bar-plot for the assessed variables in Figure 5. The most influential variable in the dataset was GDP\_million$, which had a strong positive correlation with the target variable (People Positive New Cases). The GDP\_million$ variable represented the GDP in 2019 for each country or state, a good estimator for each country’s share in the global market in 2019. This result supported the ideas drawn during the EDA phase that more frequent trading activities had a terrible impact on the spread of the epidemic. In addition, its impact on the spread of the epidemic was four times more than the REPORT\_DATE (time) did. The high positive correlation REPORT\_DATE variable and the target variable indicated that the pandemic would spread quickly over time as expected. The opposite influences that Land\_km2 and Area\_km2 did on the target were quite interesting. The Area\_km2 represented the total area of that region for both land area and water area, so this result indicated more portion of the water area in that region would positively affect the containment of the epidemic. Chart

Description automatically generated

*Figure 5*. Distribution of feature importance.

The future trends of the spread of the epidemic for each state in the United States were forecasted using the generalized autoregressive conditional heteroskedasticity (GARCH) model. This model was chosen for its ability to describe and forecast the volatility in time-series analysis. The GARCH model utilized the data and residues obtained from the previous training to assist the variance of logged data to forecast the future volatility of the target variable, making it an ideal choice for forecasting the future volatility of the epidemic spread. There were 50 models in total, each for a different state in the United States, as during the model training phase, each state data was trained separately. The primary purpose of the modeling was to forecast the positive new cases for each state in the next 90 days from April 2. The approach was to use the first 348 sample data to fit the model and used the data of the last 90 days as validation. The average score of the models was around 2000; this indicated an overall relatively good fit for all models. The result interactive plot was available in the published final project website, demonstrating the forecasted volatility of new positive cases for each state.

**RECOMMENDATIONS**

The spread patterns of the COVID-19 epidemic forecasts could be further improved by studying each country’s epidemic data at a more granular geographical level. In this way, it is possible to analyze the interaction between the epidemics in each small area by learning the response of each small area to the spread of the epidemic. Additionally, studying the transmission of previous epidemics and summarizing the characteristics of each epidemic, then comparing the direct commonalities and transmission of each epidemic, could help to generalize the pattern of epidemic transmission and provide valuable experiences for humankind in fighting the future epidemic.

**CONCLUSION**

RNA viruses have appeared more and more frequently since the 21st century, and their powerful infectiousness can spread among various spinal species. This characteristic makes it difficult for every RNA virus to be completely controlled. Even if the virus is contained within a human range, the spread between animals is difficult to control. This characteristic of the RNA virus leaves a significant hidden danger for the second spread and recurrence of the virus. The most effective solution is a vaccine, but a new vaccine requires much clinical verification. This fact indicates that in the long period of the virus outbreak, the only way to fight the epidemic is to control the epidemic. COVID-19 is a typical example. It took a year for the vaccine to be successfully developed. During this time, human society suffered huge losses. Over 153 million people were infected, and more than 3 million people lost their lives. However, all of this can be reduced by studying the pattern of the spread of the epidemic. Through the learned patterns, the government can allocate medical resources and implement epidemic prevention measures more efficiently, such as when and where to implement stricter lockdowns. More detailed data can significantly improve the efficiency of learning. To achieve this requires a more integrated epidemic surveillance system, the correct guidance of governments worldwide, and cooperation among the people.

**BIOGRAPHY**

**Junhe Zhang** is a graduate student in the Data Science Program at The George Washington University. His interests include artificial intelligence, deep learning, natural language processing, and stock market data analysis. He enjoys doing meaningful and challenging work and loves world peace.

**Dr. Nima Zahadat** is a professor of data science, information systems security, and digital forensics. His research focuses on studying the Internet of Things, data mining, information visualization, mobile security, security policy management, and memory forensics. He has been teaching since 2001 and has developed and taught over 100 topics. Dr. Zahadat has also been a consultant with federal government agencies, the U.S. Air Force, Navy, Marines, and the Coast Guard. He enjoys teaching, biking, reading, and writing.

**REFERENCES**

Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digit. Med.* **4,**3 (2021). <https://doi.org/10.1038/s41746-020-00372-6>

Liu, Antonio. “Letter to the Editor: Exponential Increase in COVID-19 Related Publications Compared to Other Pandemic Diseases.” Rambam Maimonides medical journal vol. 12,1 e0009. January 19 2021, doi:10.5041/RMMJ.10424

J Woolhouse, Mark E et al. “RNA Viruses: A Case Study of the Biology of Emerging Infectious Diseases.” Microbiology spectrum vol. 1,1 (2013): 10.1128/microbiolspec.OH-0001- 2012. doi:10.1128/microbiolspec.OH-0001-2012

Srivastava, N., Baxi, P., Ratho, R. K., &amp; Saxena, S. K. (2020). Global trends in epidemiology of Coronavirus DISEASE 2019 (COVID-19). Medical Virology: From Pathogenesis to Disease Control, 9-21. doi:10.1007/978-981-15-4814-7\_2

Gallaway MS, Rigler J, Robinson S, et al. Trends in COVID-19 Incidence After Implementation of Mitigation Measures — Arizona, January 22–August 7, 2020. MMWR Morb Mortal Wkly Rep 2020;69:1460–1463. DOI: <http://dx.doi.org/10.15585/mmwr.mm6940e3>

Srivastava, N., Baxi, P., Ratho, R. K., & Saxena, S. K. (2020). Global trends in epidemiology of Coronavirus DISEASE 2019 (COVID-19). *Medical Virology: From Pathogenesis to Disease Control,* 9-21. doi:10.1007/978-981-15-4814-7\_2

Mohan BS, Nambiar V (2020) COVID-19: An Insight into SARS-CoV-2 Pandemic Originated at Wuhan City in Hubei Province of China. J Infect Dis Epidemiol 6:146. doi.org/10.23937/2474-3658/1510146

**APPENDIX A**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Type** | **Description** |
| GDP\_million$ | Int | GDP in 2019 |
| REPORT\_DATE | Str | Report data for each data point |
| Land\_km2 | Int | Land Area in km2 |
| Avg(age) | Int | Average age in that region |
| Population | Int | Number of populations in that region |
| Latitude | Int | Region coordinate data |
| Longitude | Int | Region coordinate data |
| Area\_km2 | Int | Total area including land and water in km2 |