International Core Journal of Engineering Volume 6 Issue 11, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

Novel Coronavirus 2019—Visualization and Prediction

Zhaoxuan Lai[1], Benjamin Liang[2], Chuwei Xu[3]

[1]Jinan-Birmingham joint institute, Jinan University, Guangzhou 511443, China;

[2]Shanghai American School PD, Shanghai 200135, China;

[3]Tandon School of Engineering, New York University, New York City 11201, USA.

Abstract

The Novel Coronavirus 2019 is a crucial issue that has been a major global concern over the recent months. Visualizing its current situation and predicting its future development is beneficial for people to accurately recognize the circumstance and to take effective measures accordingly. This work uses novel charts to visualize Covid-19 on the global scale and continental scale with correlation analysis among and spread analysis. K-means clustering is also applied to group countries by number of infections and linking them with the countries' unique policies in combating the virus. An innovative method to predict the short evolution future of the virus is provided using a regression with Leaky ReLu. Shelford's law of tolerance is referred to and compared with. The visualization suggests that North America has the greatest number of confirmed cases, and the situation is worsening over time. A positive relation between the numbers for confirmed, deaths, and recovered cases is observed. The approximate burst time and spread route of the virus in different continents can also be derived. The country policy analysis compares the measures taken by various countries around the world with its results and concludes that more rigorous measures are generally more effective at controlling the spread of the virus. The future evolution trend of the virus is dependent on the behaviors of people is implied from the prediction analysis. This work aims to help people gain a better understanding of the current situation and the future evolution of Covid-19 as well as providing insights for improving measures adopted by countries in combating the virus.

Keywords

Novel Coronavirus 2019; Visualization; Prediction.

1. Introduction

1.1 Our work

This work uses the up-to-date Novel Coronavirus 2019 data [1] to create an analysis through Python. The analysis is made in the form of visualization and prediction. In total there are three sets of data: global time series data of confirmed, recovered, and deaths numbers, each consisting of five attributes. In the visualization part, the analysis was made both on the continental scale and for a few specific countries combined with their government's policy in fighting against this epidemic. While in the prediction part, only the results of each continent are studied separately.

The main libraries applied in the visualization part are Pyecharts, Plotly and Matplotlib. Pyecharts is top notch at making various types of visualizations and Plotly is widely used to make interactive graphs. Matplotlib is used to perform K-means clustering. In the prediction part, the Leaky ReLu model of Keras is applied for the global spread prediction.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

1.2 Data Preprocessing

The data of this work consists of csv files which provide very comprehensive oversight on the Covid-19 situation, with time series data tracking daily updates on Confirmed Cases, Deaths, and Recovered Cases, all of which are marked with their respective countries, continents, and latitude and longitude values. Because of how encompassing and well organized the data is, it is able to provide a solid basis for performing analyses and for making use of the daily update statistics in the visualizations.

This work uses the pycountry_convert library, which provided some conversion functions, and some of the countries in the data were renamed afterwards in order to fit into the format required. This work organized the data into countries and continents, and removed the unnecessary latitude and longitude values to complete the data preprocessing.

2. Visualization

2.1 Overview

The visualization component began by displaying the overview statistics of worldwide coronavirus cases. This was demonstrated through a global heatmap that was created using the Plotly library. It follows a red color scheme and indicates the severity of the coronavirus(determined based on total confirmed cases in a country) using shades and tints. The heatmap format allows for comparisons between countries to be easily observed and for viewers to quickly discern between regions that are impacted more and less by the coronavirus. Through the visualization, it can be seen that the United States and Brazil are among the countries with the highest confirmed cases.

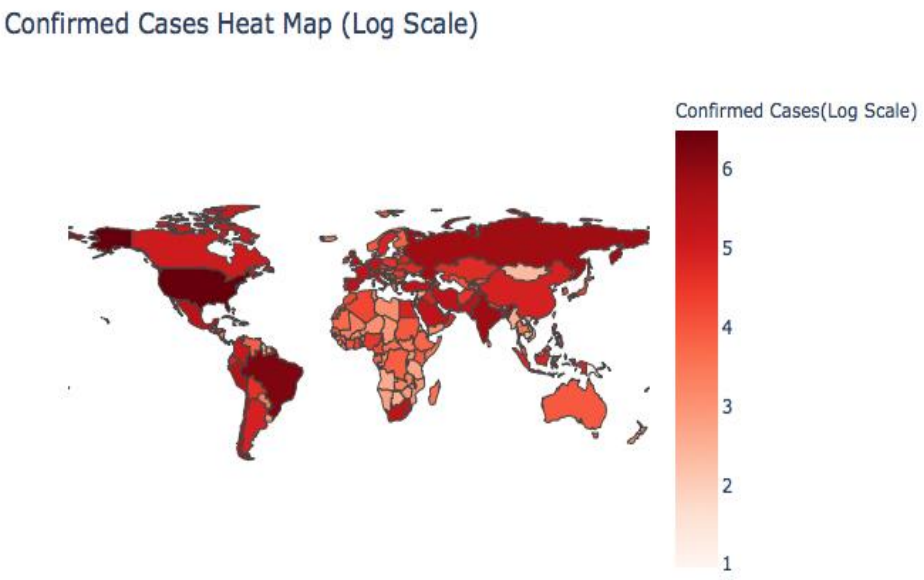

Figure 1. Heatmap of confirmed cases worldwide

This work continued by categorizing the visualization by continent. The library used to create this diagram is Pyecharts. A stacked area polar chart was chosen as it can display the confirmed cases, recovered cases, and deaths all together while ensuring that comparisons can be easily made between the various continents. Another benefit of this diagram is that the ratio of the three aforementioned values for each continent can be perceived through this setup. Due to the large disparity between the number of deaths and the number of confirmed cases for each continent and how they follow the same y-axis, the sector representing deaths is very small and barely visible for each continent. In the chart, it can be observed that the values for Australia and Others are significantly lower than the other continents with its total confirmed cases being even lower than the recovered cases of other continents such as Asia and Europe. This is illustrated in **Figure 2**.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

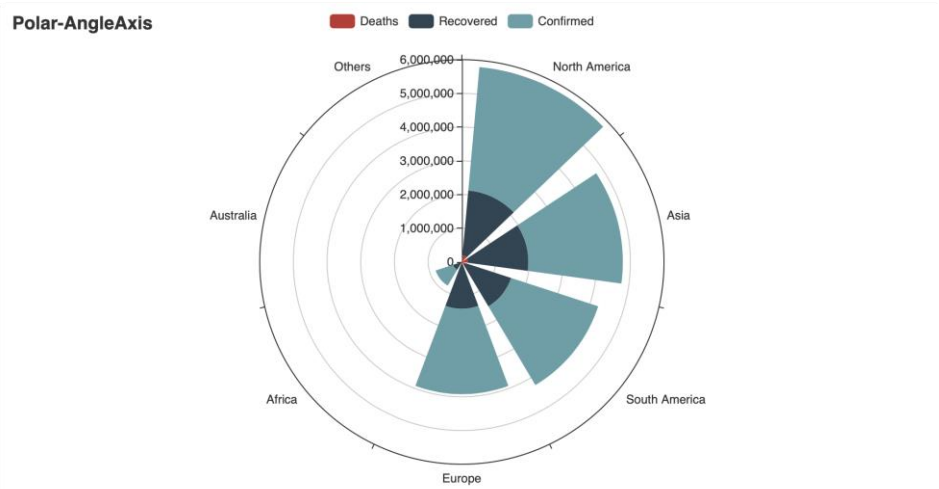

Figure 2. Polar chart of confirmed cases, recovered cases, and deaths by continent

2.2 Correlation analysis

A 3D scatter plot is used to demonstrate the correlation between the total numbers of confirmed, recovered and deaths. Pyecharts is the main library used to draw this graph. Each point represents the total number of confirmed, recovered and death cases in that continent. Due to the relative low numbers of total cases in Australia and others, their locations on the plot are approximately around zero. As seen through the representations of North America, Asia, South America and Europe, there is always a positive correlation between any two of the three variables, and these continents are relatively high in all three values.

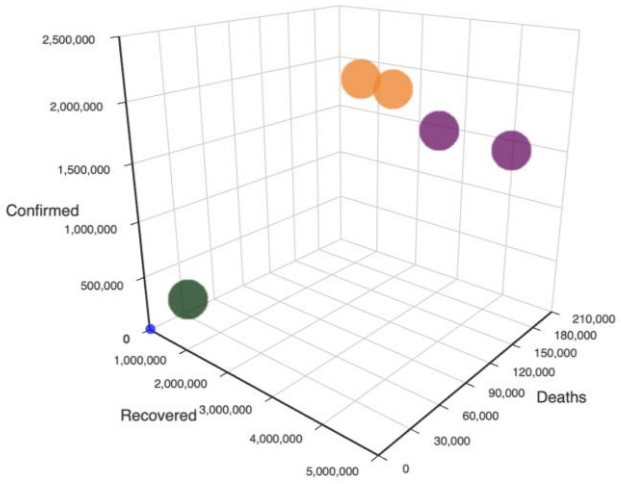

Figure 3. 3D scatter plot of the total number of confirmed, recovered and death cases in each continent

2.3 Spread analysis

The spread analysis is made using Plotly to visualize the spread progression of the coronavirus on the world map as well as create a 2D scatter plot of the change in death rates and confirmed cases in each continent. Both of which are interactive graphs that the time-line can be dragged to view the spread progression of confirmed cases and death rates in each continent. **Figure 4** and **Figure 5** illustrate the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

situation up to July 8, 2020. From **Figure 4,** the outbreak sites in different time periods and the approximate spread route of the virus globally can be observed.

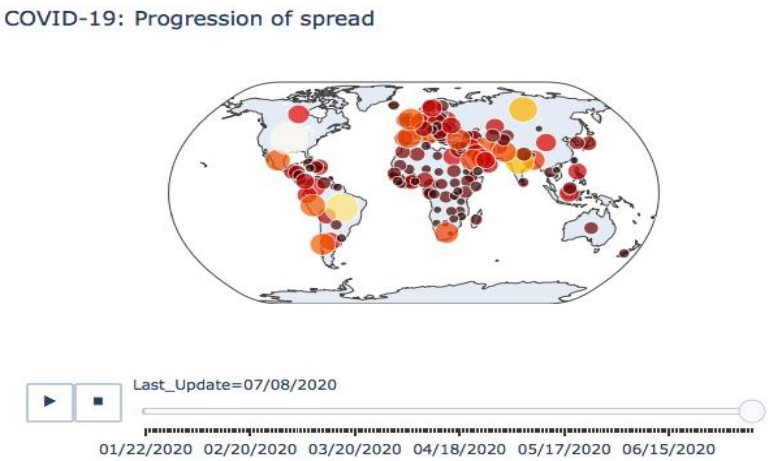

Figure 4. Progression of spread of the Novel Coronavirus 2019 on world map

Figure 5 illustrates the results of tracking confirmed cases and death rates in each country and classified by different continents. The result implies the spread trend in different continents and through the comparisons of which, the distribution range of death rates and confirmed cases for each continent are perceivable. Also, seen from the graph below, there is a positive linear relation between confirmed cases and death rate.

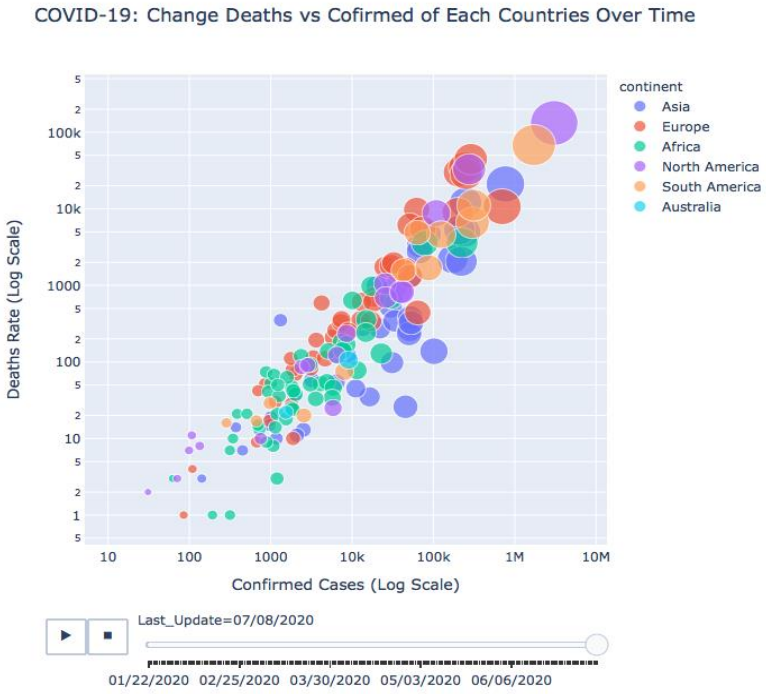

Figure 5. 2D scatter plot of change in death rates and confirmed cases in each continent

Pyecharts is used to make the following stacked area curve(**Figure 6**) and line plot(**Figure 7**). The stacked area curve demonstrates the change in total confirmed cases in each continent over time. Since the number of total confirmed cases in Australia and Others are insignificant compared to the remaining continents, their growth line is close to the horizontal axis. Rest of the continents all have an upward trend, among which North America, Asia and South America are increasing more rapidly

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

while Europe and Africa increase smoothly. Over time, the growth rate is slowing down for Europe and Africa. However, for North America, Asia and South America, the total confirmed cases are rising at an increasing speed. The difference in total confirmed cases among continents is becoming larger and larger as time goes by. Furthermore, it can be observed that on average, the outbreak of the virus started in April and evolved into a serious global issue afterwards.

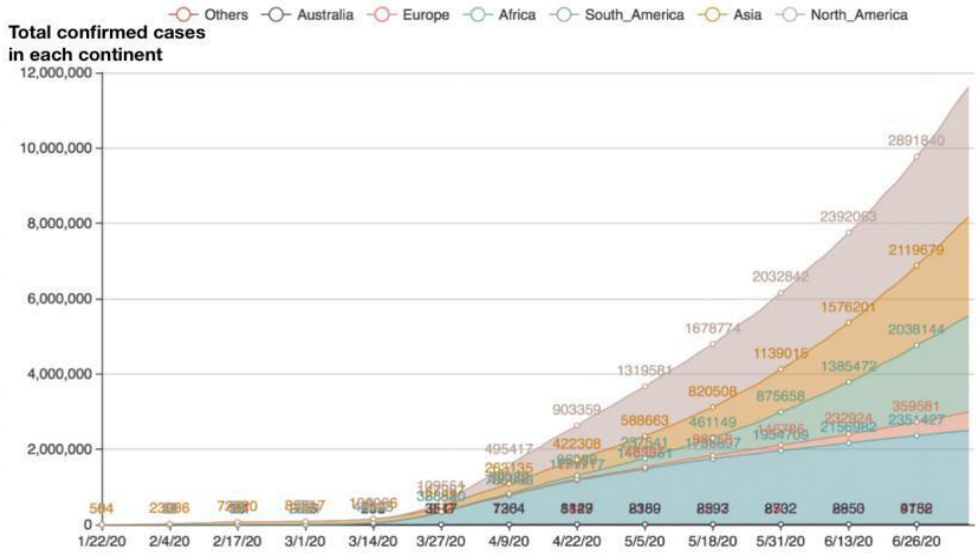

Figure 6. Stacked area curve of change in confirmed cases in each continent

The daily new confirmed numbers are calculated in order to illustrate the increase more directly and concretely.

$$\boldsymbol{Daily\ new\ confirmed\ cases_t = Total\ confirmed\ cases_t - Total\ confirmed\ cases_{t-1}} \quad (1)$$

The following line plot (**Figure 7**) shows the change in daily new confirmed cases in each continent. From which it can be seen that North America, Asia, and South America have a similar trend, where the daily new confirmed cases are quite volatile but still in a rising trend. Europe and Africa have a similar trend, which moves more steadily. Under this comparison, **Figure 7** indicates that the daily new confirmed cases in North America, Asia and South America are still increasing sharply, while the daily new confirmed cases in Europe and Africa increase less dramatically.

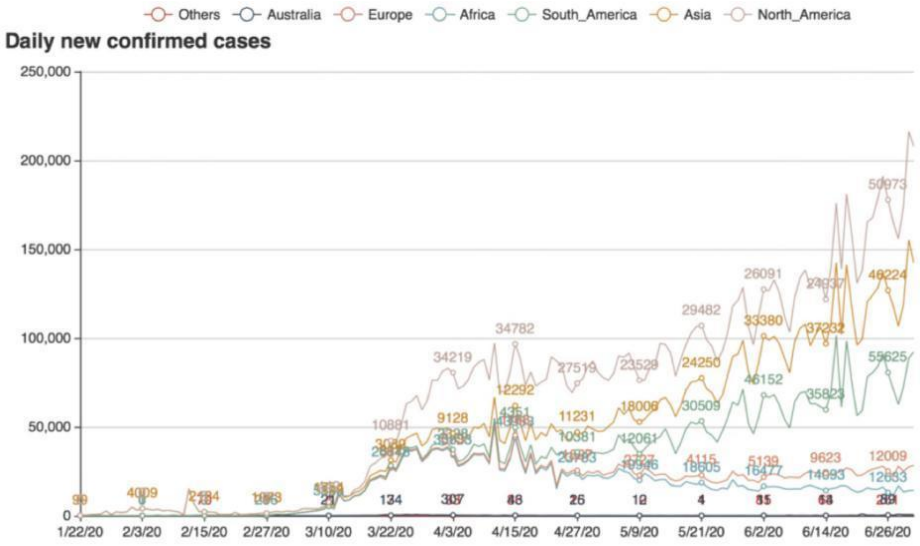

Figure 7. Line plot of change in daily new confirmed cases in each continent

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

The above analysis induces that in the early period of the epidemic, from January to March, the virus broke out in Asia initially. In April, the virus began to ravage the entire world. Currently, the situation is improving in Europe and Africa while it is worsening in North America, Asia and South America. Among those, North America is experiencing the worst impact.

2.4 Case study: specific country analysis based on their policies

To gain a better understanding of the present situation in each country, K-means clustering is applied to group each country based on their respective confirmed and death cases. Given that the USA and Brazil have much more confirmed and death cases compared to other countries, they are in a cluster separate from all other countries. India and Russia also have noticeably high numbers and belong to one cluster. What leads to the significant difference in cases between countries is closely related to their established government policies to combat the virus.

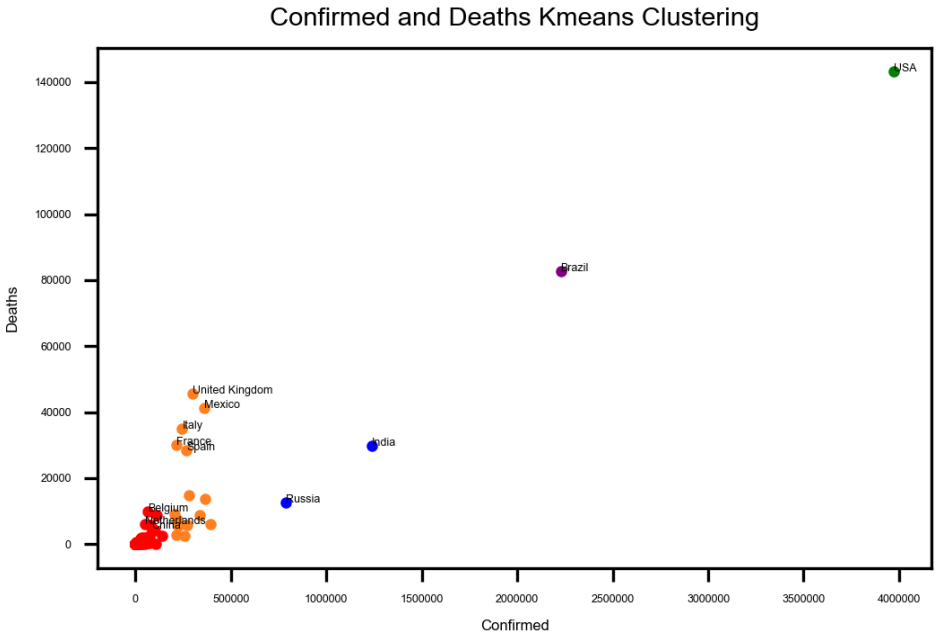

Figure 8. K-means Clustering of confirmed and deaths cases in each country

2.4.1 The USA and China

Figure 9. Protesters gather for the “Unmask Us” protest in Scottsdale, USA [4]

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

China is the country with the earliest outbreak of the virus. However, up to July 2020, China is in the cluster of colour red with relatively low numbers of confirmed and death cases, while the USA is in the green cluster with the highest numbers of both confirmed cases and deaths in the world. This large disparity results from their varied responses to the virus. First, in the early stage of the epidemic, China decisively took measures to close down the cities, which, to a certain extent, curbed the outbreak of the virus in China. However, the USA did not make the greatest use of the best epidemic prevention period by learning from the experience of China. When they started to take action later, the situation had spread to an extent that was very difficult to contain. Second, during the epidemic, Chinese people resolutely carried out the virus prevention work by following the government's appeal to wear masks and not gathering together. [2] On the contrary, many Americans do not wear masks, as many believe that wearing masks hinders their right to breathe freely. Also, they do not consider wearing masks to be necessary and helpful in preventing the spread of the virus. Third, many Americans were initially unsupportive of the home quarantine order. The U.S. government is largely at fault for the situation as it does not recognize the severity of the epidemic and advocate enough to its citizens. [3]

2.4.2 France and Spain

From **Figure 8**, France and Spain are in the same cluster, marked by the colour orange. Both countries have strict policies that restrict the citizens to stay at home as much as possible. Spain was once one of the countries most affected by the virus. At the peak of the crisis, Spain reported nearly 10,000 newly confirmed cases in one day. Since then, the country, guided by strong surveillance, testing, contact tracing, treatment and isolation, has made considerable efforts to curb transmission. The shift resulted from the strong determination of the Spanish public to comply with tight restrictions, including blockades, physical distance and other important measures to control transmission. Together, these efforts have successfully changed the direction of the epidemic in the country. [5]

2.4.3 Other countries

According to the latest data, the USA, Brazil, India and Russia accounted for over 50% of the total confirmed cases globally (up to the end of July 2020). [6] Passive and layback policy is the main reason for the surge of confirmed cases in recent time periods. Rigorous policies should be enforced accordingly in order to prevent the situation from getting worse.

3. Global trend prediction

Predictions are made based on past data and events. All the information about the spread of the virus is concluded and a pattern is found, where the change of the spread in the short future constructed from past and current data can be inferred, however the outcome that results from data that will be generated in the future cannot be provided.

Based on the data obtained from Github, the trends are able to be analyzed on global scale and continental scale. This prediction is only based on the current spread trend, and since no other factor is included, this trend may be quite different from reality. Also, the testing capacity will become the limiting factor, so it may affect the actual figure a lot. This is to estimate the possible values that could be reached with the consideration of any measures taken by different countries to control the spread. The Artificial Neural Network is applied to construct the prediction model. Leaky Rectified Linear Unit (Leaky ReLu) is chosen as the activation function, which has the mathematical formula of the form:

$$f(x) = x^{+} = max\,(0, x) \tag{2}$$

Leaky ReLu is a preferred activation function that looks and acts like a linear function, however is indeed a nonlinear function allowing complex relationships in the data to be learned. It provides more sensitivity to the activation sum input and avoids easy saturation as well. [7]

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

From **Figure 10**, the slope of the line is decreasing, which means that as time goes by, the growth rate will become approximately zero as expected. From **Figure 11,** the prediction model using the Leaky ReLu shows an exponential trend. Comparing it with Shelford's law of tolerance, as shown in **Figure 12,** which states that an organism's success is based on a complex set of conditions and that each organism has a certain minimum, maximum, and optimum environmental factor or combination of factors that determine success [8], the future evolution pattern can be approximately predicted. Though a virus cannot be considered as an organism, the Shelford's law of tolerance could be a decent reference.

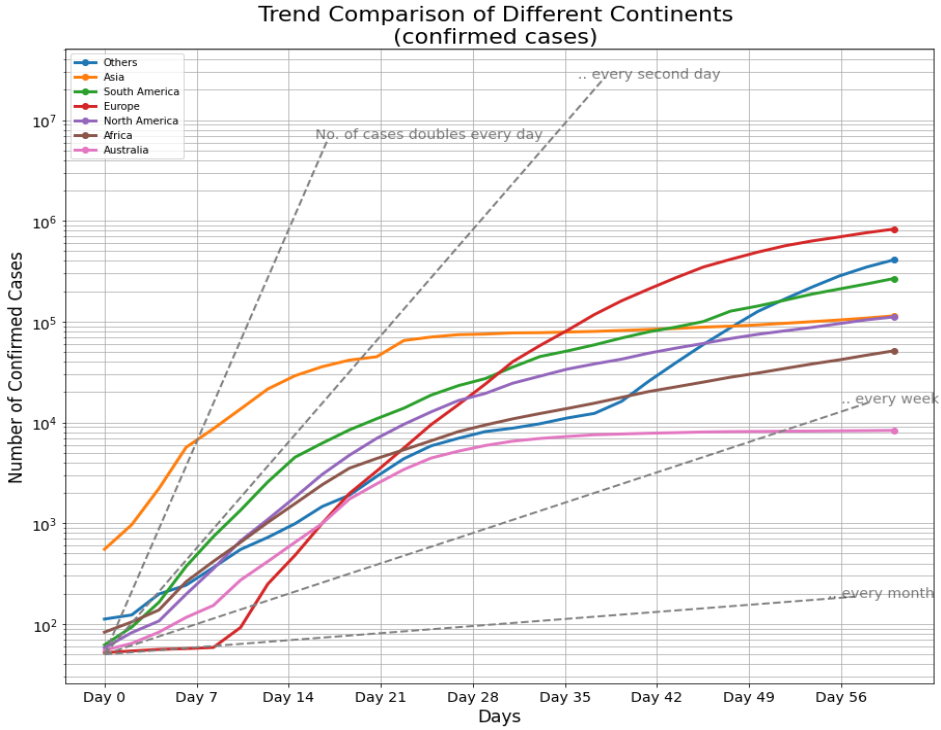

Figure 10. Spread trend of Covid-19 in different continents

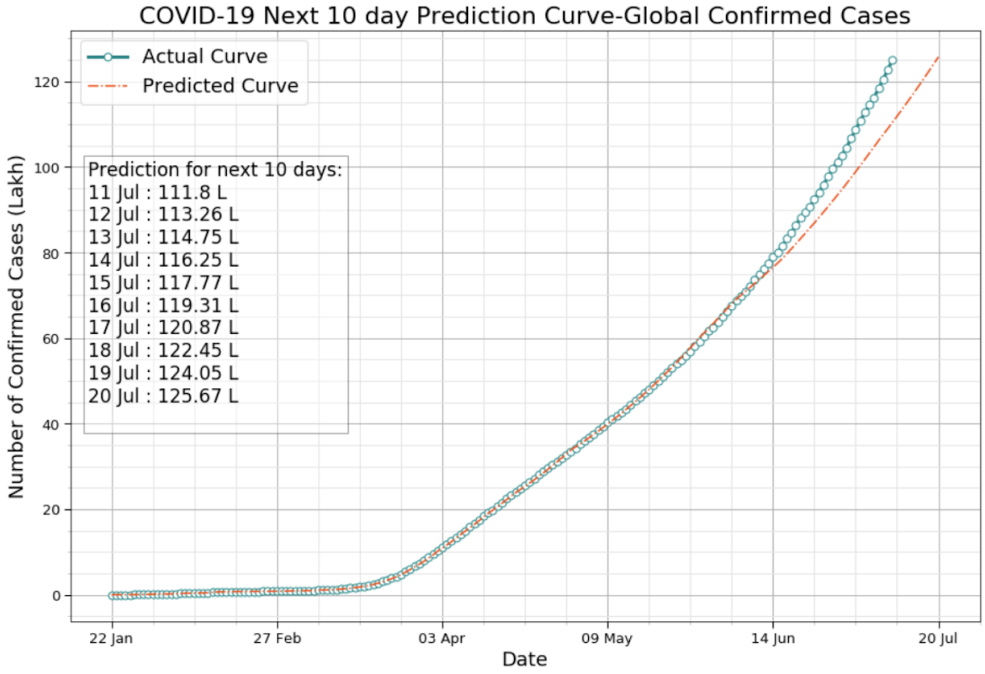

Figure 11. Prediction model of confirmed cases for Covid-19

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

From **Figure 12**, the first part of the curve in the optimal zone is quite similar to the exponential trend we have in the spread of the virus and this part represents the burst time throughout the world. The growth rate is relatively high during that time. After the critical point, which is the highest point in the curve, the spread will have two patterns. The first pattern is staying at the highest level as the red line shows, which means that the virus will still exist and won't be exterminated. In this case, people retain their normal lifestyles and take no actions in combating the virus. The second pattern is that the curve is decreasing and going to the stress zone, which means people are washing their hands, wearing masks and keeping social distance. Under this circumstance, the virus will come into the stress zone, the population of which will decrease into a low quantity. However, the virus cannot step into the intolerance zone because of two reasons. First, the vaccine hasn't come up so the virus cannot be totally exterminated. Second, the government of each country's own strategy will also affect the trend of the spread.

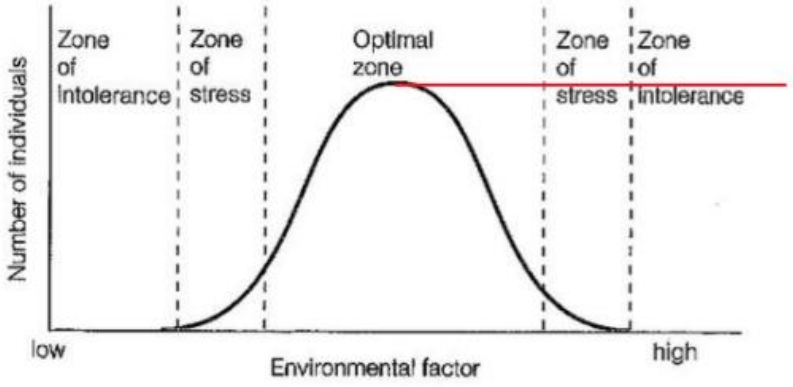

Figure 12. Shelford's law of tolerance [9]

Strategies of each country and their own situation also affect the spread of the virus. Take India as the first example. According to some research, India has issues in controlling the spread of the virus. The number of confirmed cases reported by India is quite small and the world is misled by this number. In fact, India has many hiding cases. Only their upper level classes can have the opportunity to get tested for the virus. For the poor, they are not able to protect themselves as their government does not have the ability to offer enough medical care for them.[10] Let England be the second example. The British government essentially allowed all their citizens to get infected and announced that all of them would be immune to this virus after being infected, also known as herd immunity. [11] This policy has resulted in high infection numbers. Other countries have attempted to curb the spread of the virus and keep infection cases as low as possible. For instance, the Chinese government has come up with some policies to help their citizens survive under this virus. People are asked to stay at home and keep social distance. Also, they are advised to wash hands frequently and wear masks when going outside. The strict enforcement of these policies has resulted in a relatively low amount of cases in the country.

4. Conclusion

This work studies the Novel Coronavirus 2019 through the analysis of visualization and prediction. The visualization provides a global overview of Covid-19, a correlation analysis, a spread analysis, as well as a discussion of the specific policies of various countries in combating the virus. The visualization provides a clear demonstration of the current situation of Covid-19 and its impact on people's life. In the prediction part, with the application of the Leaky ReLu, the future spread trend of the virus is predicted. Also, compared with the Shelford's law of tolerance, the future evolution pattern of this epidemic has been analyzed. Seen from the visualization, North America has the greatest confirmed cases and the situation of which is getting worse. A positive relation between the numbers for confirmed, deaths, and recovered cases is observed. The approximate burst time and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0015

spread route of the virus in different continents can also be derived. Concluded from the countries' policy analysis, for several countries in severe epidemic, more rigorous measures should be enforced in order to prevent the situation from getting worse. From the prediction, the future evolution trend of the virus depends on the behaviors of people. This work aims to help people gain a better understanding of the current situation and the future development of Covid-19 as well as providing insights for improving measures adopted by countries in combating the virus. Later studies may include the latest measures undertaken by countries and learn the future evolution pattern of the epidemic by constructing more advanced models, rather than the method derived in this work (comparing with the Shelford's law of tolerance) [12]. Various regression models should be tested in order to discover the best fit. For instance, Random Forest Regression [13], Bayesian Ridge Regression [14] and Logistic curve [15].

References

[1] The JHU data set is available on GitHub: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

[2] Xiong, Y., Mi, B., Panayi, A.C., Chen, L. and Liu, G. (2020), Wuhan: the first post-COVID -19 success story. Br J Surg. doi:10.1002/bjs.11875

[3] The Lancet. COVID-19 in the USA: a question of time. Lancet. 2020;395(10232):1229. doi:10.1016/S0140-6736(20)30863-1.

[4] Catherine Rafferty/The Republic., https://www.usatoday.com/story/news/nation/2020/07/15/arizona-most-resistant-wearing-covid-19-face-masks-study-finds/5442738002/.

[5] The resilience of the Spanish health system against the COVID-19 pandemic. Legido-Quigley, Helena et al. The Lancet Public Health, Volume 5, Issue 5, e251 - e252.

[6] Coronavirus 2019-nCoV, CSSE .Coronavirus 2019-nCoV Global Cases by Johns Hopkins CSSE. (Available from: https:// gisanddata. maps. arcgis. com/ apps/ opsdashboard/ index. html#/ bda7594740fd40299423467b48e9ecf6).

[7] Wikipedia contributors. (2020, July 31). Rectifier (neural networks). In Wikipedia, The Free Encyclopedia. Retrieved 07:13, August 2, 2020, https://en.wikipedia.org/w/index.php?title=Rectifier_(neural_networks)&oldid=970447823.

[8] Wikipedia contributors. (2020, April 23). Shelford's law of tolerance. In Wikipedia, The Free Encyclopedia. Retrieved 02:28, August 4, 2020, from https://en.wikipedia.org/w/index.php?Title=Shelford%27s_law_of_tolerance&oldid=952650399.

[9] Bioknowledgy https://www.bioknowledgy.info/c1-species-and-communities.html.

[10] Biswas, Soutik. "The 'mystery' of India's Low Covid-19 Death Rate." BBC News, 28 Apr. 2020, www.bbc. com/ news/world-asia-india-52435463.

[11] Siddique, Haroon. "The UK Government's Changing Coronavirus Strategy." The Guardian, 1 July 2020, www.theguardian.com/world/2020/may/06/the-uk-governments-changing-coronavirus-strategy.

[12] N. I. Sapankevych and R. Sankar, "Time Series Prediction Using Support Vector Machines: A Survey," in IEEE Computational Intelligence Magazine, vol. 4, no. 2, pp. 24-38, May 2009, doi: 10.1109/ MCI. 2009.932254.

[13] Kane, M.J., Price, N., Scotch, M. et al. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 15, 276 (2014). https://doi.org/ 10.1186/1471-2105-15-276.

[14] Sofiane Brahim-Belhouari, Amine Bermak, Gaussian process for nonstationary time series prediction, Computational Statistics & Data Analysis, Volume 47, Issue 4, 2004, Pages 705-712, ISSN 0167-9473, https://doi.org/10.1016/j.csda.2004.02.006.

[15] Stan Lipovetsky (2010) Double logistic curve in regression modeling, Journal of Applied Statistics, 37:11, 1785-1793, DOI: 10.1080/02664760903093633.