Statistical Analysis and Visualization of the Potential Cases of Pandemic Coronavirus

Muthusami Rathinasamy¹ and Saritha K²

¹Dr Mahalingam College of Engineering and Technology
²PA College of Engineering and Technology

May 11, 2020

Abstract

A local outbreak of initially unknown cause pneumonia was detected in Wuhan (Hubei, China) in December 2019 and a novel coronavirus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was quickly found to be causing it. Since then, the epidemic has spread to all of China’s mainland provinces as well as 58 other countries and territories, with more than 87,137 confirmed cases around the globe, including 79,968 from China, 7,169 from other countries as of 1 March 2020, as stated by the World Health Organization (WHO) in the COVID-19 situation report - 41. In response to this current public health emergency, this study done a statistical analysis and visualized reported cases of coronavirus disease 2019 (COVID-19) based on the open data collection provided by Johns Hopkins University. Where the location and number of confirmed infected cases have been shown, there have also been deaths, recovered cases and comparisons of the growth rates between the Globe countries. This was intended to provide researchers, public health officials and the general public with exposure to the epidemic.

1 INTRODUCTION

Diseases and bacteria or viruses which cause them often have different names. The "human immunodeficiency virus," HIV, for example, induces the "acquired immunodeficiency disease," AIDS. The virus that triggers the current outbreak is called coronavirus 2, a serious acute respiratory syndrome shortened to SARS-CoV-2. The illness, shortened to COVID-19, is called coronavirus disease. The World Health Organization, and the International Committee on Virus Taxonomy (https://talk.ictvonline.org/), gave these names. In public speaking, the WHO also refers to the virus as "the virus accountable for COVID-19," or "the COVID-19 virus." The outbreak was first reported in Wuhan city, China. Wuhan is the capital of the Hubei Province and has a population of around 11 million. Chinese authorities reported a cluster of related pneumonia cases in the town on 29 December 2019. A novel coronavirus which was later called SARS-CoV-2 soon confirmed to cause these cases (https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china, 2020; https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports, 2020; Ren, Wang, & Wu 2020; Drosten, Günther, & Preiser, 2003; Chen, Liu, & Guo, 2020; Zhu, Zhang, Wang, 2020). The first COVID-19 cases outside of China were found in Thailand on January 13, and in Japan on January 16 (https://www.cdc.gov/coronavirus/2019-nCoV/summary.html, 2020). The Chinese Government put the city of Wuhan and other cities in the area on lockdown on January 23rd. COVID-19 has since spread to several more countries-cases have been recorded in all regions of the world. It grew into a global pandemic by March, and was announced by the WHO as such (Yoo, 2020; https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen, 2020; Hui, I Azhar, Madani, 2020). While people sometimes refer to the virus that causes COVID-19 as "the coronavirus," several different coronaviruses do exist. The word refers to a group of viruses specific to humans: coronaviruses cause about 30 percent of all cold cases (Mesel-Lemoine, Millet, Vidalain, 2020). Corona is Latin for "crown" – this group of viruses is named because, under an electron microscope, its surface looks
like a crown. As the outbreak of the novel SARS-CoV-2 is increasingly spreading in China and beyond, threatening to become a global pandemic, epidemiological data need to be interpreted in such a way that the model of statistical data analysis and visualization can increase the understanding of situation among the mass population in the coming days (Muthusami, Bharathi, Saritha, 2020).

2 DATA SOURCE AND METHODS

2.1 Dataset

The World Health Organization (WHO), Johns Hopkins University researchers, and other agencies all maintain dataset on the number of cases, deaths, and disease recoveries. All data obtained in this research work is from Johns Hopkins University and is freely accessible via the GitHub repository. The dataset covered the period from 22 January 2020 to 17 April 2020 which includes time-series and aggregated data (https://github.com/CSSEGISandData/COVID-19, 2020).

2.2 Statistical Analysis

We statistically analyzed our dataset with various methods of data analysis and visualized those data to provide a proper understanding of the COVID-19 outbreak worldwide. Our exploit analysis was carried out by Johns Hopkins University with the 2019 coronavirus dataset (January–April 2020). Here, between 22 January 2020 and 17 April 2020, we present an effort to visualize and analyze the results. COVID-19 has so far propagated nearly 185 Countries / Regions, 83 Cities / Provinces have been registered, and 264 separate geographical locations combined. By using time-series data estimated the number of specific cases, such as confirmed infected, deaths, and recovered around the globe and the top 10 countries in the world.

Worldwide the total confirmed cases are 2,152,646, and the global average rate is 0.38. The reported cases of the world’s top 10 countries are listed below with global percentage on time-series data. The confirmed infected cases of the top 10 countries in the world are listed below with the global percentage. In this case, the US ranked first with a total of 667,801, the Global percentage is 31.02, and with a total of 184,948, the Global percentage of 8.59, Spain is second.

| Country.Region | Province.State | Totals | GlobalPerc | LastDayChange | t-2 | t-3 | t-7 | t-14 | t-30 |
|----------------|----------------|--------|------------|---------------|-----|-----|-----|------|------|
| US             |                |        | 31.02      |               | 28680 27051 31824 1362 |
| Spain          |                |        | 8.59       |               | 5103 4603 5233 1391 |
| Italy          |                |        | 7.85       |               | 2667 2972 3951 4207 |
| France         |                |        | 6.78       |               | 3217 5955 4342 5233 1391 |
| Germany        |                |        | 6.40       |               | 3394 1287 3990 6365 3070 |
| United Kingdom |                |        | 4.79       |               | 4617 4603 5252 8681 4450 676 |

Worldwide ts-confirmed Totals: 2152646
The estimated number of deaths worldwide is 143,800, with a global average of 3.61. For this situation, the US occupied the first place with 32,916 counts, and with 22,170 counts, Italy was second in the top 10 countries around the world. The total number of cases recovered is 542,107 worldwide. In this scenario, Germany ranked first, with a total of 77,000, and with a total of 74,797, Spain placed second in the top 10 countries of the world.

The aggregated data for various cases such as confirmed infected, dead, recovered and active in the top 10 countries around the world are then computed and shown below. In this scenario, the number of countries reported is 185, the number of cities / provinces listed is 138 and the total number of geographic locations combined is 3042. The aggregated data for confirmed infected cases of the top 10 countries worldwide are provided below. In this case, Spain is in the first position, with an estimate of 184,948, while Italy is in the second position, with an estimate of 168,948.

| Country/Region | Confirmed Cases | Confirmed Perc. | Deaths | Deaths Perc. | Recovered | Recovered Perc. | Active | Active Perc. |
|----------------|-----------------|-----------------|--------|--------------|-----------|-----------------|--------|--------------|
| Spain          | 184,948         | 8.59            | 19,315 | 10.44        | 74,797    | 40.44           | 90,836 | 49.11        |
| Italy          | 168,941         | 7.85            | 22,170 | 13.12        | 40,164    | 23.77           | 106,607| 63.10        |
| France         | 145,960         | 6.78            | 17,920 | 12.28        | 32,812    | 22.48           | 95,228 | 65.24        |
| Germany        | 137,698         | 6.40            | 4,052  | 3.02         | 77,000    | 55.92           | 56,646 | 41.14        |
| New York City, New York, US | 123,146 | 5.72            | 11,477 | 9.32         | 0         | 0.00            | 111,669| 90.68        |
| United Kingdom | 103,093         | 4.79            | 13,729 | 13.32        | 0         | 0.00            | 89,364 | 86.68        |
| Iran           | 77,995          | 3.62            | 4,869  | 6.24         | 52,229    | 66.96           | 20,897 | 26.79        |
| Turkey         | 74,193          | 3.45            | 16,43  | 2.21         | 70,899    | 95.5            | 65,461 | 88.23        |
| Hubei, China   | 67,803          | 3.15            | 3,222  | 4.75         | 64,435    | 95.03           | 146    | 0.22         |
In aggregate death-related results, Italy ranks first with a count of 22,170 and Spain ranks second with a count of 19,315 out of the top 10 countries in the world. In the aggregated recovered cases, Germany is in the first place with a total of 77,000, and Spain is in the second position with a total of 74,797 out of the top 10 countries in the world. In aggregate active cases, the United States (New York) ranked first with 111,669 counts and Italy ranked second with 106,607 counts from the top 10 countries around the world. The overall summary of confirmed infected, fatal and recovered cases with respect to time series in terms of totals, average and standard deviations is provided below.

******************************** OVERALL SUMMARY*****************************
**************************************************************************

**** Time Series TOTS ****
ts-confirmed ts-deaths ts-recovered
2152646 143800 542107
6.68% 25.18%

**** Time Series AVGS ****
ts-confirmed ts-deaths ts-recovered
8153.96 544.7 2168.43
6.68% 26.59%

**** Time Series SDS ****
ts-confirmed ts-deaths ts-recovered
46373.8 3069.33 9782.04
6.62% 21.09%

* Statistical estimators computed considering 250 independent reported entries

From a statistical data analysis, it can be understood that 5% of deaths and 8% of recoveries occurred in reported cases in the United States. In Spain, 10% of deaths and 40% of recoveries occurred in confirmed cases. 13% of the deaths and 24% of those recovered from Italy and 3% of the deaths and 56% of those recovered from Germany occurred in the reported case.

2.3 Visual Data Analysis

We also explore time-series data using visual data analysis to provide a clear and understandable outcome of this extreme outbreak of COVID-19. This segment will analyze various time-series data using several visual data analysis approaches with the R programming language. We have created a graph and given awareness of how SARS-CoV-2 spread around the globe from 22 January 2020 to 17 April 2020; it allows individuals to grasp the epidemiological essence of COVID-19.

Figure 1. Global confirmed infected, recovered, deaths and active cases as at 17-04-2020

Figure 1 indicates that the confirmed infected cases have been crossed by 2,000,000 cases. Many cases, such as death, recovery and active, have also been shown. New cases reported on a single day do not actually represent new cases on that day, as the number of confirmed infected cases or deaths announced by any organization – including WHO, ECDC, Johns Hopkins University and others – does not reflect the total
number of new cases or deaths on that day. This is due to the long chain of reporting that occurs between a new case or death and its inclusion in national or international statistics.

The steps in this chain vary among countries, but for many countries the reporting chain contains several of the following steps: 1. Doctor or laboratory diagnoses the case of COVID-19 on the basis of a test or combination of symptoms and epidemiological likelihood (such as a positive family member check). 2. The doctor or laboratory shall send a report to the health department of the city or district. 3. The Health Department receives a report and reports every individual case, including patient details, in the reporting system. 4. The Ministry or some other government agency gathers these data and releases the latest figures. 5. International data organizations such as the WHO or the ECDC can then compile statistics from hundreds of these national accounts.

This reporting chain will take several days to complete. This is why the numbers published at any given date do not generally reflect the number of new cases or deaths at that particular date. Confirmed deaths to date, we know the total number of confirmed deaths due to COVID-19 to date. Limited research and difficulties in the classification of the cause of death mean that the number of confirmed deaths might not be an accurate count of the actual total number of deaths from COVID-19. Death or recovery for all cases is not yet established in the current epidemic of final outcomes. The time from symptom onset to death for COVID-19 varies from 2 to 8 weeks (https://github.com/CSSEGISandData/COVID-19, 2020). It means that certain people who are already infected with COVID-19 will be killed at a later date. As discussed below, this needs to be held in mind when comparing the current number of deaths with the current number of incidents.

Regression and generalized linear models of data from the COVID-19 time series are used to analyze confirmed infected, deaths and recovered cases. The fitted models have yielded better statistical results; the graphic findings shown below represent all three cases in the USA. From the models results obtained, on the confirmed case, the exponential model coefficients are -0.807 and 0.17, the GLM Poisson model coefficients are 3.469 and 0.119, and the GLM Gamma model coefficients are -0.433 and 0.17, both of which are statistically significant, as shown in Figure 2. In case of death, the exponential model coefficients are -2.774 and 0.144, the GLM Poisson model coefficients are -2.424 and 0.151, both of which are statistically reasonable, as shown in Figure 3. In the recovered case, the exponential model coefficients are -2.204 and 0.137, the GLM Poisson model coefficients are -2.864 and 0.163, both of which are statistically significant, as shown in Figure 4.

Figures 2, 3 and 4 below display the numerous incidents, such as confirmed infected, deaths, and recovered cases in the United States.

Figure 2. US – Confirmed infected case
Figure 3. US – Death case
Figure 4. US – Recovered case

From Figures 2, 3 and 4, we can understand that all cases, such as confirmed infected, deaths and recovered, are exponentially increased, the same thing is reflected in the upper part of the graph, i.e. the output of linear and generalized linear models.

Figures 5, 6 and 7 show confirmed, fatal and recovered cases in Spain. Here, too, the count has risen exponentially; the same trend is statistically reflected in the upper part of the chart. In the confirmed case, the exponential model coefficients are -2.278 and 0.185, the GLM Poisson model coefficients are 4.159 and 0.093, both of which are statistically significant, as shown in Figure 5. In case of death, the exponential model coefficients are -2.919 and 0.152, the GLM Poisson model coefficients are 1.329 and 0.104, both of which are statistically fine, as shown in Figure 6.

Figure 5. Spain – Confirmed case
Figure 6. Spain – Deaths case
In the recovered case, the exponential model coefficients are -2.876 and 0.165, the GLM Poisson model coefficients are 0.914 and 0.124, both of which are statistically appropriate, as shown in Figure 7. In the event of an outbreak of an infectious disease, it is important not only to monitor the number of deaths, but also the pace of growth at which the number of deaths is that. If there is a fixed number of deaths over a fixed duration, we call that "linear" growth. But if they continue to double within a fixed time span, we call that "exponential" growth.

Figure 7. Spain – Recovered case

Figures 8, 9 and 10 show that the rate of growth in all cases, such as confirmed infected, deaths, and recovered, is shown in the US. Looking at the rate of death growth, we can understand that it’s exponential growth in the US.

Figure 8. US – Rate of growth in confirmed case
Figure 9. US – Rate of growth in deaths case
Figure 10. US – Rate of growth in recovered case

Figures 11, 12 and 13 show that the growth rate of all cases, such as confirmed infected, deaths, and recovered, has risen in Spain. If we look at the rate of death growth, we can understand that it is exponential growth in the US, the last day of change is 31,451 as of the April 17, 2020 study.

Figure 11. Spain – Rate of growth in confirmed case
Figure 12. Spain – Rate of growth in deaths case
Figure 13. Spain – Rate of growth in recovered case

Figure 14 indicates that changes every day occurred in confirmed cases between 23 January 2020 and 15 April 2020 from the USA and Spain. By this we will conclude that the reported cases will increase exponentially on 20 March 2020 and that the last day of change is 31,451 in the US. In Spain, the confirmed case rises linearly from 03 March 2020 to 15 April 2020, the last day of change is 7,304.

Figure 14. US vs Spain - Changes per day

Figure 15 shows that confirmed, deaths, recovered and active cases of the United States and Spain, along with global time series data for all these cases, occurred between 22 January 2020 and 17 April 2020. The chart brought more clarification to the above analysis findings, especially with regard to the United States and Spain. In the same way, our proposed research is capable of estimating the number of cases for a given country across the globe and can also compare the different cases between countries. For example, Figure 16 shows confirmed infected, deaths, recovered and active cases across the globe and throughout the United States, France, China, Spain, Germany, Italy and India. From this Figure, all cases, such as confirmed, deaths, recovered, and active deaths, can be understood to be minimal in India as at present (17 April 2020).

Figure 16. Confirmed, deaths, recovered and active cases in Globe, US, France, China, Spain, Germany, Italy, and India

It is clear that the real-time analysis of these data is extremely useful in documenting the epidemiological behavior of this severe disease. We believe that this method of data analysis will certainly boost understanding of the situation and inform behavior.

3 RESULTS AND DISCUSSION

This study examined three separate categories of data, including confirmed infected, death and recovered cases across the globe, for the period from 22 January to 17 April 2020. It will also include a comparative overview of all the cases reported in the United States and Spain. Nevertheless, we are discussing various cases internationally in order to explain the various cases identified over a particular time span. After review,
2,152,646 confirmed cases of COVID-19 occurred worldwide on 17 April 2020. In the US, where the highest count is 667,801, the global percentage is 31.02. Death cases were 143,800 across the globe (6.68%), with the US top count being 32,916 (4.93%). The cases recovered were 542,107 around the globe (25.18%) with Germany at the top of the list with a total of 77,000 cases. The visual analysis of the growth rate of confirmed infected, deaths and recovered cases between the US and Spain is another investigation.

The goal of this article on COVID-19 is to summarize existing research, collect relevant data and make it possible for readers to make sense of the published data and early research on the coronavirus outbreak. Much of our work focuses on known problems for which we can link with well-established research and evidence on COVID-19. The research presented here is based on statistical and visual data analysis methods with the aid of a dataset provided by John Hopkins University. The research was done with R Studio 1.2.5033 and R 4.0 beta versions of the Windows 10 operating system. The purpose of this article on COVID-19 is to aggregate existing research, bring together the relevant data and allow readers to make sense of the published data and early research on the coronavirus outbreak. Most of our work focuses on established problems, for which we can refer to well-established research and data on COVID-19. The analysis presented here based on statistical and visual data analysis with the help of the dataset provided by John Hopkins University. The analysis was made with R Studio 1.2.5033 and R 4.0 beta version in Windows 10 operating system. Each and every description of the different cases of COVID-19 is documented here between 22 January 2020 and 17 April 2020. We are now also observing the harmful outbreak of the SARS-CoV-2 virus. To the world, this is extremely troubling. In this analysis, we examined the top 10 countries most affected and comprehensive reported cases of the United States and Spain.

4 CONCLUSION

In conclusion, the dataset COVID-19 (2019-nCoV) from the Johns Hopkins CSSE data repository (22 January 2020 to 17 April 2020) was used for our experiment. It has supported us to generate and disseminate detailed information to the scientific community and to the public, especially at the peak phase, in order to understand the growth and impact of the novel coronavirus. Nevertheless, knowledge of this novel SARS-CoV-2 virus remains minimal among the general population around the globe. Raw data published from different sources are not adequately capable of offering an insightful understanding of COVID-19 as a consequence of SARS-CoV-2. A user-friendly data analysis platform would also be more effective in recognizing the epidemic of this severe disease. The informative and interactive graphics of the visualization framework provide an intuitive interface and a clear view of each and every raw data. Hopefully, in the coming days, we will continue to track the epidemiological data of this outbreak that we have used in this report and from other official sources.

ETHICAL STATEMENT

We declare that ethical statement is not applicable.

CONFLICT OF INTERESTS

There was no conflict of interest with others.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

2019 Novel CoronaVirus CoViD-19 (2019-nCoV) Data Repository by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) https://github.com/CSSEGISandData/COVID-19 (Accessed 17 April 2020)

Centers for Disease Control and Prevention. 2019 Novel Coronavirus (2019-nCoV), Wuhan, China. 2019. https://www.cdc.gov/coronavirus/2019-ncov/summary.html(Accessed 15 January 2020)
Chen, Y., Liu, Q., Guo, D. (2020). Emerging coronaviruses: genome structure, replication, and pathogenesis. *J Med Virol*. 92, 418-423. https://doi.org/10.1002/jmv.25681

Dey, S. K., Rahman, M. M., Siddiqi, U. R., & Howlader, A. (2020). Analyzing the epidemiological outbreak of COVID-19: a visual exploratory data analysis (EDA) approach. *J Med Virol*. 92(6), 593-607. https://doi.org/10.1002/jmv.25743

Drosten, C., Gunther, S., Preiser, W., et al. (2003). Identification of a novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*. 348, 1967-1976.

Hui, D. S., I Azhar, E., Madani, T. A., et al. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health— the latest 2019 novel coronavirus outbreak in Wuhan, China. *Int J Infect Dis*. 91, 264-266.

Mesel-Lemoine, M., Millet, J., & Vidalain, P. O., et al. (2012). A human coronavirus responsible for the common cold massively kills dendritic cells but not monocytes. *J Virol*. 86(14), 7577-7587. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3416289/. doi:10.1128/JVI.00269-12

Muthusami, R., Bharathi, A., & Saritha, K. (2020). COVID-19 Outbreak: Tweet based Analysis and Visualization towards the Influence of Coronavirus in the World. *J Gend & Org*. 33(2), 534-549.

Ren, L. L., Wang, Y. M., Wu, Z. Q., et al. (2020). Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study [published online ahead of print, 2020 Feb 11]. *Chin Med J (Engl)*. doi:10.1097/CM9.0000000000000722

World Health Organization. (2020). Naming the coronavirus disease (COVID-19) and the virus that causes it. The ICTV’s page is here: International Committee on Taxonomy of Viruses (ICTV). https://talk.ictvonline.org/ (Accessed 20 March 2020)

World Health Organization. (2020). Coronavirus disease 2019 (COVID-19) situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (Accessed 5 April 2020)

World Health Organization. (2020). WHO statement regarding cluster of pneumonia cases in Wuhan, China. Jan 9, 2020. https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china (Accessed 15 February 2020)

World Health Organization. (2020). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf (Accessed 15 March 2020)

World Health Organization. (2020). Rolling updates on coronavirus disease (COVID-19). 2020https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen (Accessed 20 March 2020)

Yoo, J.H. (2020). The fight against the 2019-nCoV outbreak: an arduous march has just begun. *J Korean Med Sci*. 35, e56. https://doi.org/10.3346/jkms.2020.35.e56

Zhu, N., Zhang, D., Wang, W., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 382, 727-733. https://doi.org/10.1056/NEJMoa2001017

Figure 1. Global confirmed infected, recovered, deaths and active cases as at 17-04-2020

Figure 2. US – Confirmed infected case

Figure 3. US – Death case

Figure 4. US – Recovered case

Figure 5. Spain – Confirmed case
Figure 6. Spain – Deaths case
Figure 7. Spain – Recovered case
Figure 8. US – Rate of growth in confirmed case
Figure 9. US – Rate of growth in deaths case
Figure 10. US – Rate of growth in recovered case
Figure 11. Spain – Rate of growth in confirmed case
Figure 12. Spain – Rate of growth in deaths case
Figure 13. Spain – Rate of growth in recovered case
Figure 14. US vs Spain - Changes per day
Figure 15. Confirmed, deaths, recovered and active cases in Globe, US and Spain
Figure 16. Confirmed, deaths, recovered and active cases in Globe, US, France, China, Spain, Germany, Italy, and India
Figure 2. US – Confirmed infected case

![Graph showing confirmed infected cases in the US.]

Figure 3. US – Death case

![Graph showing death cases in the US.]

Figure 4. US – Recovered case

![Graph showing recovered cases in the US.]

Note: These figures illustrate the trends in confirmed infections, deaths, and recoveries in the US. The graphs are based on modeled data with certain parameters and assumptions.
Figure 5. Spain – Confirmed case

Figure 6. Spain – Deaths case
Figure 7. Spain – Recovered case

Figure 8. US – Rate of growth in confirmed case
Figure 9. US – Rate of growth in deaths case

Figure 10. US – Rate of growth in recovered case
Figure 11. Spain – Rate of growth in confirmed case

Figure 12. Spain – Rate of growth in deaths case
Figure 13. Spain – Rate of growth in recovered case

Figure 14. US vs Spain - Changes per day
Figure 15. Confirmed, deaths, recovered and active cases in Globe, US and Spain
Figure 16. Confirmed, deaths, recovered and active cases in Globe, US, France, China, Spain, Germany, Italy, and India