VirusDis. (April–June 2020) 31(2):204–208
https://doi.org/10.1007/s13337-020-00610-1

SHORT COMMUNICATION

Statistical analysis and visualization of the potential cases of pandemic coronavirus

R. Muthusami[1] · **K. Saritha**[2]

Received: 24 April 2020 / Accepted: 26 May 2020 / Published online: 10 June 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, 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, 7169 from other countries as of 1 March 2020, as stated by the World Health Organization 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.

Keywords COVID-19 · Statistical analysis · Visual data analysis · Coronavirus

✉ R. Muthusami
r.muthusami@gmail.com

[1] Department of Computer Applications, Dr. Mahalingam College of Engineering and Technology, Anna University, Coimbatore, Tamil Nadu, India

[2] Department of Mathematics, P.A. College of Engineering and Technology, Anna University, Coimbatore, Tamil Nadu, India

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 Taxonomy of Viruses (ICTV) [1], 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 [2–7]. The first COVID-19 cases outside of China were found in Thailand on January 13 and in Japan on January 16 [8]. The Chinese Government put the city of Wuhan and other cities in the area on lockdown on January 23. 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 2020, and was announced by the WHO as such [9–11]. 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 [12]. 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 [13, 14].

The World Health Organization (WHO), Johns Hopkins University researchers, and other agencies all maintain dataset on the number of confirmed infected 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 [15].

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. Using time-series data, it estimated the number of individual cases, such as confirmed infected, deaths and recovered around the globe and the top 10 countries in the world. As of 17 April 2020, the United States and Spain are among the top ten countries in the world. Further to the discussion on different cases, such as confirmed illnesses, deaths and recovery in those countries as seen in the Fig. 1.

Worldwide the total confirmed infected cases are 2,152,646, and the global average rate is 0.38, with a standard deviation of 2.15. The global average rate of the Top 10 countries is 7.73, with a standard deviation of 8.49. In this circumstance, 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. 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 in the world. In this scenario, Germany ranked first, with a total of 77,000, with a total of 74,797, Spain ranked second, and the US ranked fourth, with a total of 54,703, in the top 10 countries of the world.

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.

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 indicates that the confirmed infected cases have been crossed by 2,000,000 cases around the globe. 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.

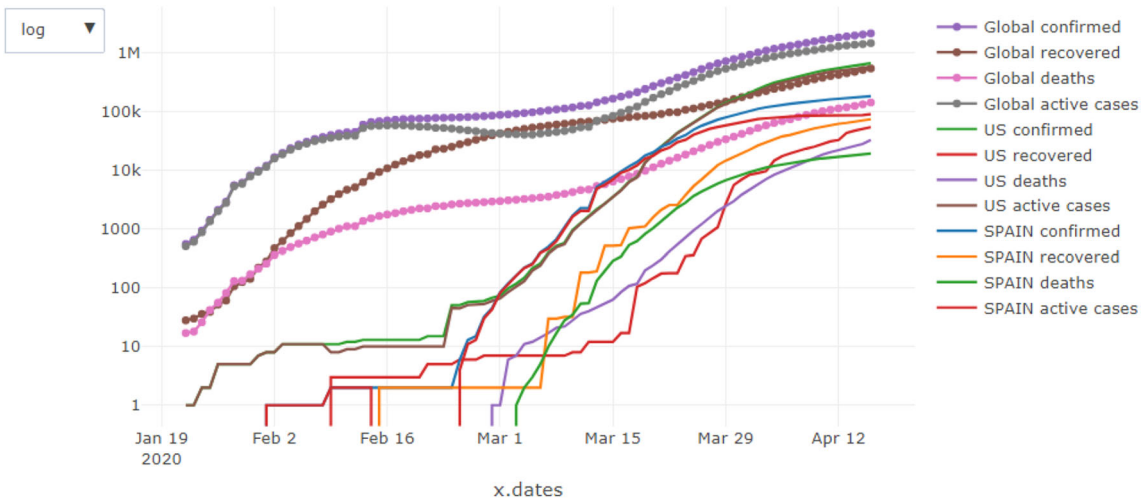

Fig. 1 Display the number of confirmed infected, deaths, recovered and active cases of Globe, USA and Spain from 22 January 2020 to 17 April 2020

Regression and generalized linear models (GLM) 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 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 Fig. 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. 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. From the findings, it can be understood that all cases, such as confirmed infected, deaths and recovered, are linearly increased, the same thing is reflected in the upper part of the graph, i.e. the output of linear and generalized linear models. Regression and generalized linear models of the COVID-19 time series were also used to analyze confirmed infected, death and recovered cases in Spain. The fitting models have provided better statistical results; the findings shown below reflect all three cases in Spain. Here, too, the count has risen linearly. 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. 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. 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.

In the event of an outbreak of an infectious disease, it is necessary not only to track the number of deaths, but also the rate of increase in the number of deaths. If there is a fixed number of deaths over a fixed period of time, we call that "linear" growth. But if they continue to double within a fixed time span, we call it "exponential" growth. Based on the results, looking at the rate of death growth, we have understood that it is linear growth in the US and Spain. Figure 3 indicates that changes every day occurred in confirmed cases between 22 January 2020 and 17 April 2020 from the USA and Spain. By this we will conclude that the reported cases will accelerate 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. 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 increase understanding of the situation and inform behavior.

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

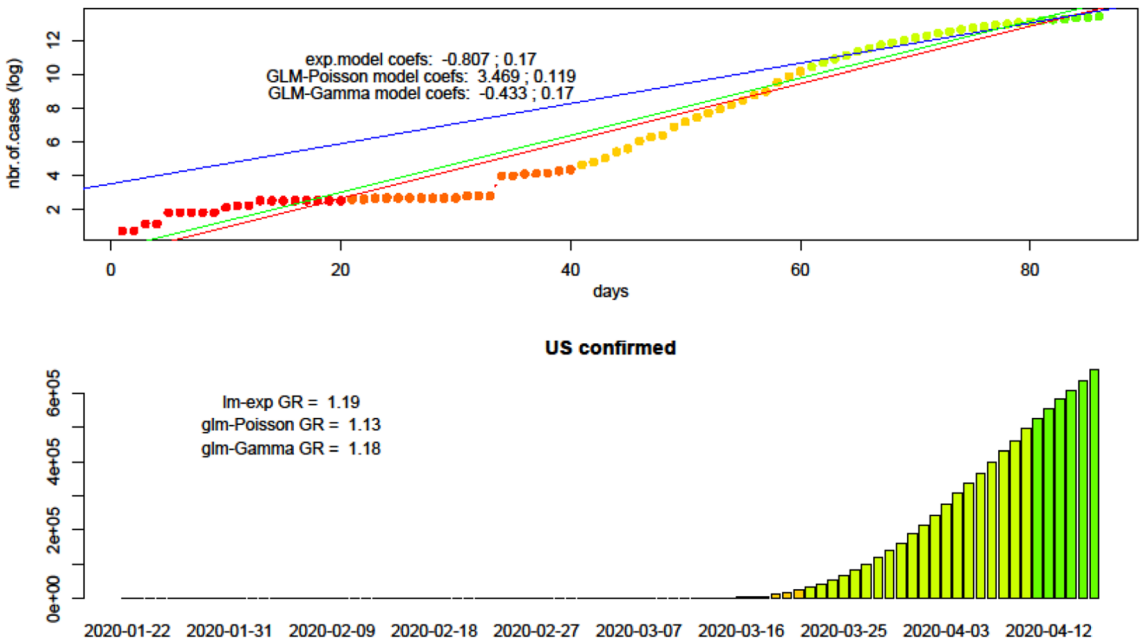

Fig. 2 Curve fitting and distribution of linear regression, and GLM with "Poisson" and "Gamma" function models with parameters of number of days and number of confirmed infections in the US between 22 January 2020 and 17 April 2020

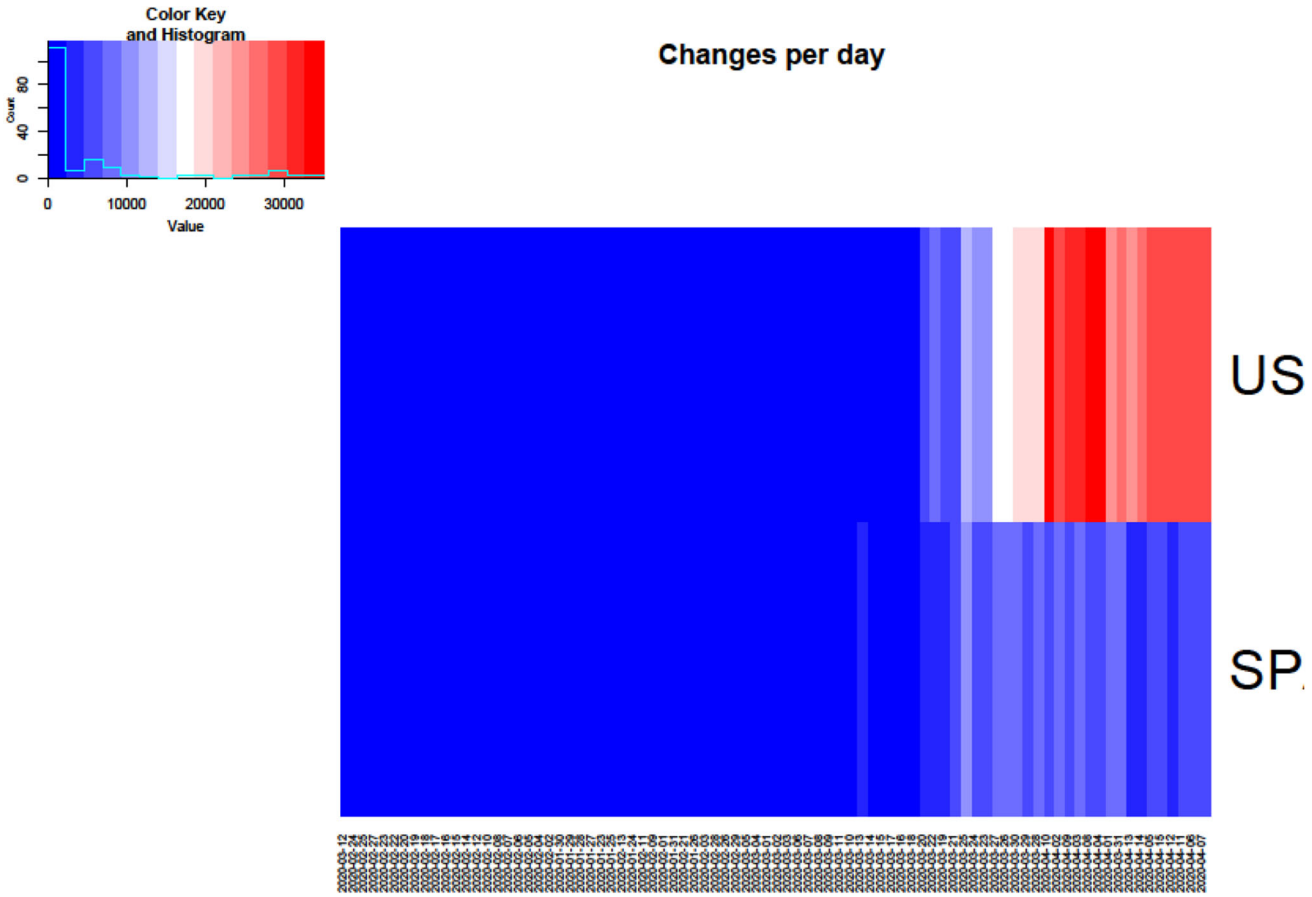

Fig. 3 Showing daily changes in confirmed cases of infection in the United States and Spain from 22 January 2020 to 17 April 2020

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. 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.

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 graphics of the visualization platform provide an intuitive

interface and a simple view of all raw data. Hopefully, in the coming days, we will continue to track the epidemiological data of this outbreak that we have used in this study and from other official sources.

Compliance with ethical standards

Conflict of interest All authors declare no conflict of interest.

Data availability statement Data will be available upon request.

References

1. World Health Organization. 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). 2020. https://talk.ictvonline.org/. Accessed 20 March 2020.
2. World Health Organization. WHO statement regarding cluster of pneumonia cases in Wuhan, Chinam Jan 9, 2020. 2020. https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china. Accessed 15 Feb 2020.
3. WHO. Coronavirus disease 2019 (COVID-19) situation reports. 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. Accessed 5 April 2020.
4. Ren LL, Wang YM, Wu ZQ, et al. 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). 2020. https://doi.org/10.1097/CM9.0000000000000722.
5. Drosten C, Günther S, Preiser W, et al. Identification of a novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med. 2003;348:1967–76.
6. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol. 2020;92:418–23. https://doi.org/10.1002/jmv.25681.
7. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382:727–33. https://doi.org/10.1056/NEJMoa2001017.
8. 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 Jan 2020.
9. Yoo JH. The fight against the 2019-nCoV outbreak: an arduous march has just begun. J Korean Med Sci. 2020;35:e56. https://doi.org/10.3346/jkms.2020.35.e56.
10. World Health Organization. Rolling updates on coronavirus disease (COVID-19). 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen. Accessed 20 Mar 2020.
11. Hui DS, Azhar EI, Madani TA, et al. 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. 2020;91:264–6.
12. Mesel-Lemoine M, Millet J, Vidalain PO, et al. A human coronavirus responsible for the common cold massively kills dendritic cells but not monocytes. J Virol. 2012;86(14):7577–87. https://doi.org/10.1128/JVI.00269-12.
13. Dey SK, Rahman MM, Siddiqi UR, Howlader A. Analyzing the epidemiological outbreak of COVID-19: a visual exploratory data analysis (EDA) approach. J Med Virol. 2020;92(6):632–8. https://doi.org/10.1002/jmv.25743.
14. Muthusami R, Bharathi A, Saritha K. COVID-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the World. J GED Organ. 2020;33(2):534–49.
15. 2019 Novel CoronaVirus CoViD-19 (2019-nCoV) Data Repository by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). 2020. https://github.com/CSSEGISandData/COVID-19. Accessed 17 April 2020.