Cloud Computing and Data Science
http://ojs.wiserpub.com/index.php/CCDS/

Case Study

Analysis and Prediction of COVID-19 Using Growth Analysis Models: A Case Study

Kalimullah Lone*, Shabir Ahmad Sofi

Department of Information Technology, National Institute of Technology Srinagar, Srinagar, India
Email: kalimullah_03phd18@nitsri.ac.in

Received: 6 December 2023; **Revised:** 21 February 2024; **Accepted:** 29 February 2024

Abstract: The coronavirus disease 2019 outbreak has added to the development of novel methods to study the epidemiological and predictive nature of the pandemic. Mining such data is necessary as this data is full of trends and information. Using data mining techniques allows us to extract and process such data to predict the pandemic's trends and behavior. Analysis, evaluation, and prediction are performed on Jammu and Kashmir's data during the period 09^{th} of March 2020 to 10^{th} of February 2021. The work is done on the dataset of patients provided by the Department of Information and Public Relations, Government of Jammu and Kashmir. Various mathematical models and techniques were used to predict the Virus spread and occurrence with the help of symptoms. We aim to propose a model to predict the virus occurrence based on the symptoms and epidemiological nature of the pandemic. The purpose of this study is to understand the virus occurrence and distribution. The work has helped our government to find the most infected areas and future challenges to tackle any such pandemic. The trends and behavior of the virus in Jammu and Kashmir were studied. People under observation, people tested for the virus, positive, negative, recovered, active, and deaths were keenly observed. The prediction to find an infected patient was carried out with the help of symptoms. The results obtained from the prediction model are verified with the actual results.

***Keywords*:** associative modelling, gompertz modelling, COVID-19, prediction

1. Introduction

Coronavirus disease 2019 spread globally with an exponential rate in a short span of time. It has forced population to live in isolation and miserable conditions. The starting stage was totally frustrating for health experts as the spreading mechanism was unknown and unusual [1]. World Health Organization (WHO) has declared corona virus disease (COVID-19) as a pandemic due to its exponential growth [2]. On 30th of January 2020, WHO called it, "public health emergency of international concern" and global threat [3]. To restrict the spreading of this pandemic, researchers are trying to figure out the nature of the virus and the factors responsible for its spread [4]. There are various factors which play a main role in the transmission of this virus like Immunity of the host, density of the population, climate or weather conditions and many more. The transmission takes place directly or indirectly from one person to other with a physical contact [5]. It is also transmitted through open air while talking, sneezing and coughing [6]. The guidelines issued by health experts and WHO to restrict community transmission are strict lock down, frequent hand wash, maintaining social distancing and many more [7]. The community transmission of this deadly virus depends upon various variables,

DOI: https://doi.org/10.37256/ccds.5220244059

which include density of the population, poor hygiene, interactions through traveling, weak public health policies and protocols [8]. There are other variables like outside temperature and seasonal weather conditions which play a role in its transmission [9]. The pandemic spread during the winter in almost 130 cities of China except Wuhan. The pandemic spread from 20th January to 2nd March 2020 in china [10].

The initial symptom of the disease was pneumonia of unfamiliar and uncommon behaviour [11]. Later on, the symptoms of the disease went on increasing day by day with an increase in the number of cases. The symptoms included fever, dry cough and fatigue and in some cases muscle soreness was also observed [12]. The initial attempts to curb the transmission of the disease was an effective quarantine. It showed a decrease in the number of cases in China and South Korea [13]. Effective quarantine yielded better results but transmission went on, when the positive case a with the family and friend circle. If an undetected case happens to be within the rescuing team, the social contacts increased to higher levels with reference to previous [14]. The quarantine formula was a wise decision to limit the pandemic but was not a cure [15]. Moreover the quarantine and isolated life has increased the stress levels [16].

The key contributions of the paper are:

1. The model has addressed all the parameters of the virus spread.

2. The model tries to find out relation between various parameters and bunches most likely parameters on the basis of combined behaviour. The allows us to find the parameters, which have an impact on the outcome of the problem.

3. The predicted growth is studied and the results obtained from step 1 and 2 are compared with that of the predicted one.

The rest of the manuscript is organized as follows: Section 2 highlights the related work and the literature studied in the same problem domain. Section 3 outlines the epidemic, associative and Gompertz modeling. Section 4 and Section 5 explain the motivation and problem formulation. Section 6 gives the proposed model and a detailed algorithm. Sections 7 and 8 deal with the case study, its results and discussion. Section 9 concludes the manuscript.

2. Related work

The number of patients increased in a short span of time, initially with a few symptoms to many and from symptomatic to asymptomatic [17]. This irregular behaviour of the virus was analysed thoroughly both medically and mathematically [18-19]. Analysis of various dataset in [20] showed recovery and death rates. The author [20] analysed a time series dataset and applied statistical methods to find the death and recovery rate. The authors in [21] worked on various datasets country wise to find the outbreak. They came with a visual exploratory analysis of died, recovered and confirmed cases. The authors in [22] used a mathematical analysis and prediction technique to find the number of susceptible, exposed, infected and recovered cases. The authors in [23] presented a prediction model based on regression analysis. The model predicted the rate of spread of COVID-19. The authors in [18] analysed various data sets through machine leaning methods like random forest and Cox survival analysis. The study has analysed deeply motility rates of various age groups. The factors taken into consideration were hazard ratio for mortality according to the geographical location of the hospitals and pandemic wave. The mortality rate is studied in every age group, in patients with lung diseases, smoking, obesity, cancer, heart failures, diabetes and many more. The authors in [24] have proposed a semi-parametric framework based on Bayesian inference. The framework combines various data sources to estimate a reproduction number, number of infected and proportion of samples with undetected cases.

The authors in [17] have discussed the role of temperature on the transmission of COVID-19. The pandemic has shown a different behaviour in different countries with the increase in temperature. In most of the countries, 5 to 8 Degree Celsius is an optimal temperature for the survival and spread of the virus [3]. The same survival and spread took place between 13 to 19 degree Celsius in some other countries [9]. The spread and survival of the virus was was found in the range 4 to 11 degree Celsius in USA and 8 to 11 in Spain [3]. There is an inverse relationship in many cases, where the spread and survival of the virus was found to be at higher temperatures [25]. In Barcelona the spread and survival took place around 38 degree Celsius [26]. An exponential growth took place between March and July in the states of Maharashtra and Punjab in India, where the temperature is above 30 Degrees [27]. The exponential growth of COVID-19 declined a bit in Kerala with the rise in temperature [28]. The authors in [3] concluded results with the remarks that COVID-19 is temperature independent, although it began in colder season.

The authors in [29] performed regression, scatterplots, distributed non linear models and random effect meta analysis to find the relationship between COVID-19 cases and temperature. The cases were collected between January 20 and February 29, 2020 in China. The samples showed that temperature and number of cases are dependant. The authors in [30] found that there was a linear relationship between COVID-19 and temperature below 3 degree Celsius and flat above 3 degree Celsius during the period 23rd January and 29th February, 2020. There is no confirmation that the cases will decrease when the weather becomes warmer. The authors in [31] strongly rejected the hypothesis that COVID-19 is temperature dependant as was previously suggested by the data sources across the world and some provinces of China. The most common respiratory problems during winter are common cold, sore throat, hoarseness, barking cough, sinus infection, Acute bronchitis, Acute exacerbations of chronic obstructive pulmonary disease (AE-COPD) and pneumonia. Analysing historical data suggest that exposure to cold through a region with low environmental temperature or hypothermia, promote high of respiratory infections both in lower and upper tracts [32]. These respiratory diseases are most common in Jammu and Kashmir where mercury rarely crosses zero degree Celsius [33]. Authors in [34] observed that the seasonal influenza viruses are common in patients with acute respiratory diseases in Jammu and Kashmir.

3. Theoretical background

In this section we will explain basic modelling techniques used in this paper. We firstly processed the available COVID-19 data to understand the trends with epidemiological models. Later, we used Apriori algorithm to predict the disease with the help of symptoms. Finally, we used Gompertz model to study the growth of the pandemic.

3.1 *Epidemic modelling*

Infectious disease modelling finds the mechanism of virus spread and its progress in the population can be shown in the Figure 1. Such models are described through various parameters. The parameters group population (N_i) into susceptible (S_j, a group not infected yet), infected (I_k, a group infected with the disease and contribute to its spread) and recovered (R_l, a group recovered after infection).

Where S_j, I_k, $R_l \subseteq N_i$.

$$s_j = S_j / N_i$$

$$i_k = I_k / N_i$$

$$r_l = R_l / N_i$$

The rate of susceptible cases or transmission with respect to time is defined as:

$$ds_j / dt = -\omega s_j i_k \tag{1}$$

where ω is the rate of contact between susceptible and infected at time t.

Rate of infected cases with respect to time:

$$di_k / dt = \omega s_j i_k - \delta i_k \tag{2}$$

where δ is the recovery rate (The value of δ can't be negative).

Rate of recovery:

$$dr_l / dt = \delta i_k \tag{3}$$

Mortality rate also comes into play when the disease becomes deadly. Thus we have mortality rate as (γi). The

mortality rate will change the rate of infected cases with respect to time as:

$$di_k / dt = \omega s_j i_k - \delta i_k - \gamma i_k \tag{4}$$

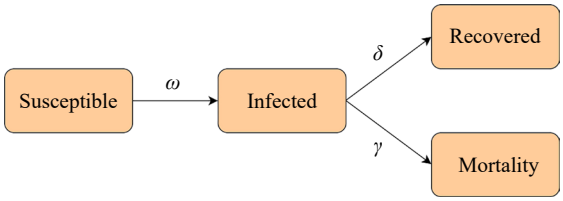


Figure 1. Virus progress in a population

Based on the above equations we can find the number of infected cases at any time. This method is based on the tests conducted in the laboratory. One we have actual values of the infected population, we will be able to predict future infected population. This will allow us to take all the precautions and medical treatments for the entire population. Expected duration of infection is i / δ. An epidemic occurs whenever the number of infected cases are on a rise, $di / dt > 0$.

This implies that

$$\omega, s_j, i_k\text{-}\delta i_k > 0, \omega, sj / \delta > 1.$$

We have $s_j = S_j / N_i$ when the susceptible cases increase exponentially and become almost equal to $N_i (S_j \approx N_i)$.

Then $\omega / \delta > 1$ because $s_j = 1$,

$$R_o = \omega / \delta > 1.$$

Where R_o is the reproduction factor.

3.2 *Associative modelling*

Associative modelling is a method to find relationship among a group of objects and variables. The associative rules find some hidden patterns and models form the existing data. The outcome of this kind of modelling is the effective and scientific decision making. The rules in associative modelling are as follows:

- We have a set of items, $I = i_1, i_2, ..., i_n$;
- The set I contains a set of transactions D which represent each transaction as t;
- X is the set of items $\in I$ and $t \supseteq I$;
- $X \subseteq I$, $Y \subseteq I$ and also $X \cap Y = \phi \Rightarrow X \rightarrow Y$;
- Confidence of $X \rightarrow Y$ is Con, if the confidence percentage of transactions $\in D$, provided $D \supseteq X$ and $D \supseteq X$;
- Support of $X \rightarrow Y$ is Sup, if the support percentage of transactions $\in D$, provided $D \supseteq X \cup Y$;

We find a set of associations from the set of transactions D. The number of associations that have greater support and confidence than the user defined minimum support and confidence are selected. We divide association rules into two steps.

Itemsets with transactions support greater than minimum support are selected.

$\forall$ itemsets, for individual itemset p, we have to find non empty subsets of p for every such subset q, the association rule is $q \rightarrow p\text{-}q$ if $\frac{support(p)}{support(q)} >$ minimum confidence.

3.3 *Gompertz modelling*

The growth of the disease is exponential and logarithm of the population size can be plotted against time as:

$$y = \ln[N / N_o]$$

Gompertz defines growth curve as a function of three parameters: one, μ_m is the maximum growth rate which actually is the tangent to the inflection point, second, λ is the lag time which is the x-intercept of the tangent and third, asymptote [$A = \ln(N_\infty / N_o)$], is the maximum value reached. The mathematical form of Gompertz equation is:

$$y = a \times e^{-e^{b-ct}} \tag{5}$$

For biological problems the values of a, b and c are replaced by A, μ_m and λ. Thus Gompertz equation takes the following form:

$$y = A \times e^{-e^{\mu_m - \lambda t}} \tag{6}$$

The inflection point of the curve is obtained by finding the second derivative of the Gompertz equation with respect to time. The second derivative at $t = t_i$, is zero, represents the inflection point. Gompertz equation takes the modifies shape as:

$$y = A \times e^{-e^{\left[\frac{\mu_m e}{A}(\lambda - t)+1\right]}} \tag{7}$$

The whole process of modeling is explained in Algorithm 1.

4. Motivation

We have discussed various problems in related work while dealing with COVID-19. From symptomatic to asymptomatic, the pandemic changed its behavior. Although the change was not so large in terms of numbers but posed a threat on the method of transmission. We will discuss only symptomatic cases because asymptomatic cases are only detected through testing (Community or volunteer testing). Symptomatic cases have defined symptoms before they are tested and detected. Based on the symptoms and seasonal diseases, it is difficult to recognize a COVID-19-positive patient. It also shows that the number of symptomatic cases is in bulk in comparison to asymptomatic cases. The patients suffering from COVID-19 follow certain stages like susceptible, positive, super spreader, under treatment, recovered, and mortality. The various stages of the pandemic are explained in the Figure 2:

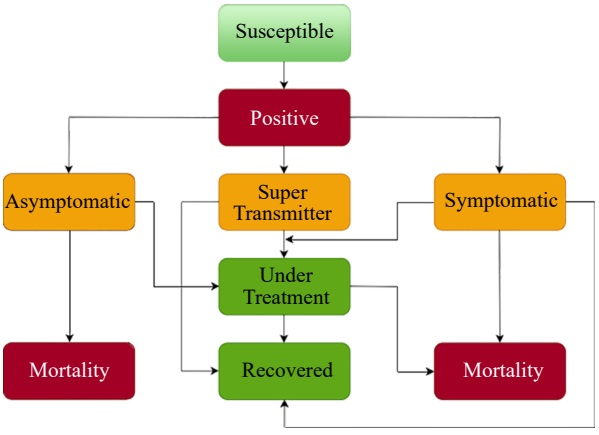


Figure 2. Pandemic stages

These diseases are found in these cold regions every year. There are chances that an individual has symptoms

due to the topology of the region but may be treated susceptible and kept in observation. Such a person can have interactions with the infected patients and later will get infected. Among all the symptoms of the coronavirus disease found in patients, there are most of the symptoms are found every year due to seasonal flu, respiratory diseases, cough, pneumonia, and fever. Due to these symptoms either found in a patient or many can never be susceptible nor infected. We have to understand the topography of the region and the previous history of all these diseases individually or two in a group or many in a group. This past experience will allow us to find the exact number of cases. Moreover, patients with such seasonal diseases are curable through medication.

5. Problem formulation

The whole world has suffered a lot due to the COVID-19 pandemic across length and breadth. The erratic behavior of this infection made it more complex and the symptoms varied from person to person. This variation in the number of symptoms from patient to patient twisted its epidemic trends after an uncertain and limitless period of time. Our aim is to use epidemic modeling (like Susceptible Infected Recovered (SIR) or Susceptible Exposed Infectious Removed (SEIR)) to find the fluctuating trends during the pandemic. The symptoms of the pandemic are studied and the apriori algorithm is used to predict, whether a patient is positive or negative based on the symptoms found. The cases predicted through apriori are verified through Gompertz model and are cross-checked with SIR model findings.

Let (N_i) be the population with susceptible (S_j), infected (I_k) and recovered $(R_l) \subseteq (N_i)$. ω, δ and γ define the rate of contact between S_j and I_k, recovery and mortality respectively at time $t' > t$. The aim of our work is to predict whether a patient is COVID-19 positive or not, based on the symptoms found in a patient. Moreover, we will find the epidemic, recoveries, and deaths from the population N_i and will predict the growth trends of the pandemic.

6. Proposed work

The proposed model is a combination of two blocks. The first block uses SIR model to find the epidemiological nature of the pandemic through various datasets. The second block predicts the future trends of the pandemic using datasets. The model bunches the symptoms together and finds the most frequent symptoms with the help of association rules. The most effective method for finding associations is the apriori algorithm. The method implies that if an item or a symptom occurs, then other items or symptoms also occur with a certain probability. The algorithm provides a symptom list of frequently occurring symptoms from a dataset. The predicted frequent itemsets make it possible for us to find a COVID-19 positive or negative case based on the symptoms of the patient. The predicted positive or negative cases are plotted using Gompertz growth curves. The output of the Gompertz is cross verified with the values of SIR model. The complete process of prediction is explained in Figure 3:

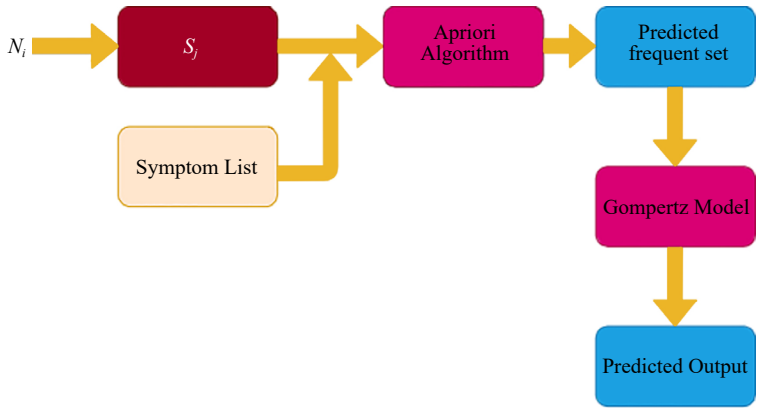


Figure 3. Prediction mechanism

7. Case study: COVID-19 outbreak in Jammu and Kashmir

The implementation is performed through PySpark and its standard libraries. The data analysis and evaluation is performed on Jammu and Kashmir's data during the period 09^{th} of March 2020 to 10^{th} of February 2021. The journey of COVID-19 in Jammu and Kashmir is summarised in Table 1. The case study focuses on the prediction based on the symptoms found in a patient. The dataset generated from the trained model is compared with the actual dataset values. The main focus is on the symptoms and differentiation between coronavirus patients and other patients with only flu, fever, cough, respiratory diseases, fatigue, etc.

The graphs show that the number of patients with asymptomatic with reference to symptomatic is much less. Moreover, the patient does not even know about the infection. So we have considered only patients with symptoms. In the initial stage, there are numerous susceptible cases, but with the passage of time, there are more and more infected patients. Due to the guidelines and Standard Operating Procedures (SOPs) issued by the government and health experts, the number of susceptible cases decreased with time. Proper Medical supervision and following SOPs increased the number of recoveries. Infected patients were on the rise initially due to a lack of knowledge of the pandemic.

Table 1. COVID-19 progress and preventive measures in Jammu and Kashmir

S No.	Date	Description
1	01/03/2020	Corona virus control team formulated
2	05/02/2020	Schools to follow SOP's strictly
3	07/03/2020	Suspension of classwork
4	09/03/2020	First Corona virus case in Jammu
5	13/03/2020	Closure of all shopping malls, Gyms, swimming pools etc
6	17/03/2020	First Corona virus case in Kashmir
7	18/03/2020	Inter state bus service stopped
8	20/03/2020	No school teacher to attend duties
9	22/03/2020	Complete shutdown except essential services
10	08/04/2020	Face mask compulsory in markets and for all on-duty officials
11	19/05/2020	Districts marked as red, orange and green zones with respect to the spread and number of cases

7.1 *Dataset characteristics*

The data is collected from the official website and Twitter handle of the Department of Information and Public Relations, Government of Jammu and Kashmir [35]. In addition to this, we have also collected sample data from the Ministry of Health and Family Welfare [36], Worldometer [37], and Kaggle [38]. The datasets provide details of COVID-19 evolution, spread, recoveries, and deaths moreover major symptoms of the disease are also discussed. The data is collected from the official government records, which cover day-by-day patient information. There is a huge list of symptoms found in the infected patients but we have merged many and restructured them into seven only. The Dataset contains 22,135,755 entries or rows. Each row contains information about the patient. There are at least 40 attributes associated with a patient. We call these attributes as symptoms of the patient. An important challenge is to reduce the number of symptoms and to check the inter-dependencies between symptoms. The symptoms which have the least impact on the outcome can be neglected. Thus, there must be a proper mechanism to verify such relations and to reduce the columns of the dataset, to make it less complex because the fewer the number of attributes will make the dataset easier to analyze.

8. Results and discussion

We analyzed data provided by the Department of Information and Public Relations, Government of Jammu and Kashmir. The analysis conducted consists of an epidemiological study of COVID-19 in a specific region (Jammu and Kashmir) and a prediction of the disease based on the symptoms.

The epidemiological study collected samples from various datasets. The samples identified different parameters like patients under observation or susceptible, total tests conducted, total samples tested negative, total samples tested positive, total recoveries, total active cases and total deaths. These parameters were observed on a daily basis. The patients have various symptoms as listed in Table 2. The symptoms in each patient are represented through symbols. The associations of these symptoms are predicted through the apriori algorithm in Figure 4. Apriori suggests that symptoms of ABC are strongly associated. Thus the occurrence of these symptoms in a patient individually or together suggests that the patient is affected by the disease. The predictions are plotted graphically through Gompertz model and compared with SIR method. The parameters were plotted with respect to date in Figure 5, Figure 7, Figure 9, Figure 11, Figure 13, and Figure 15. The curves are obtained using SIR model to find the epidemiological nature of the data. The curves Figure 6, Figure 8, Figure 10, Figure 12, Figure 14, and Figure 16 are the prediction outcomes of the Gompertz model.

Table 2. Symptoms

S No.	Symptoms	Code
1	Fever	A
2	Fatigue	B
3	Dry cough	C
4	Loss of appetite	D
5	Body ache	E
6	Shortness of breathe	F
7	Mucus	G

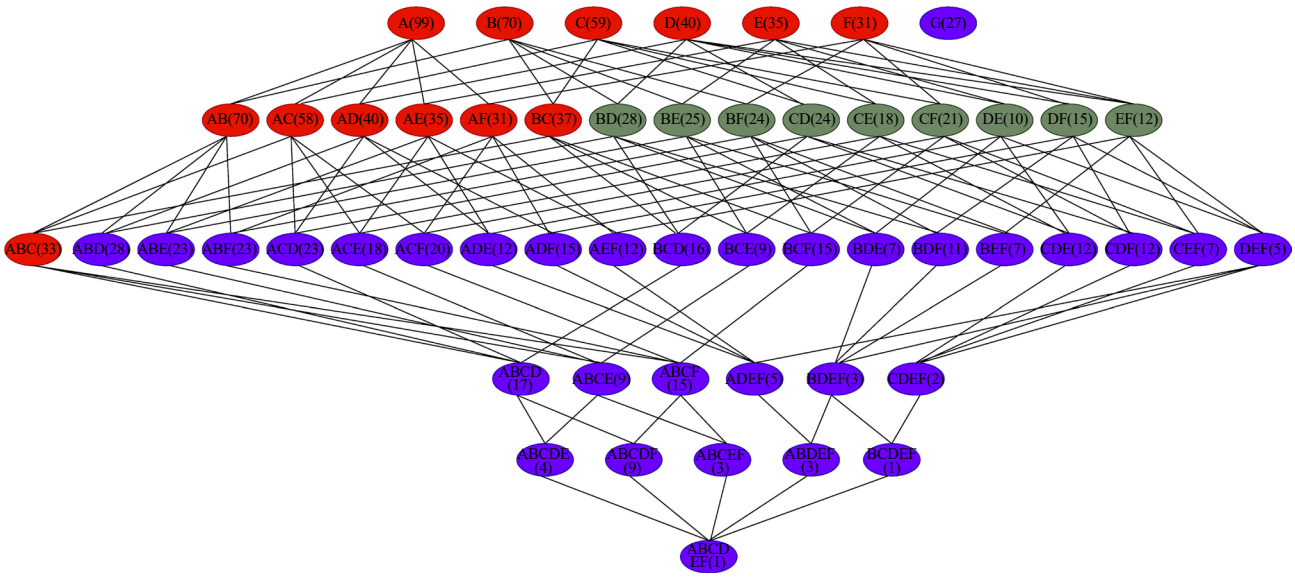


Figure 4. Frequent itemsets

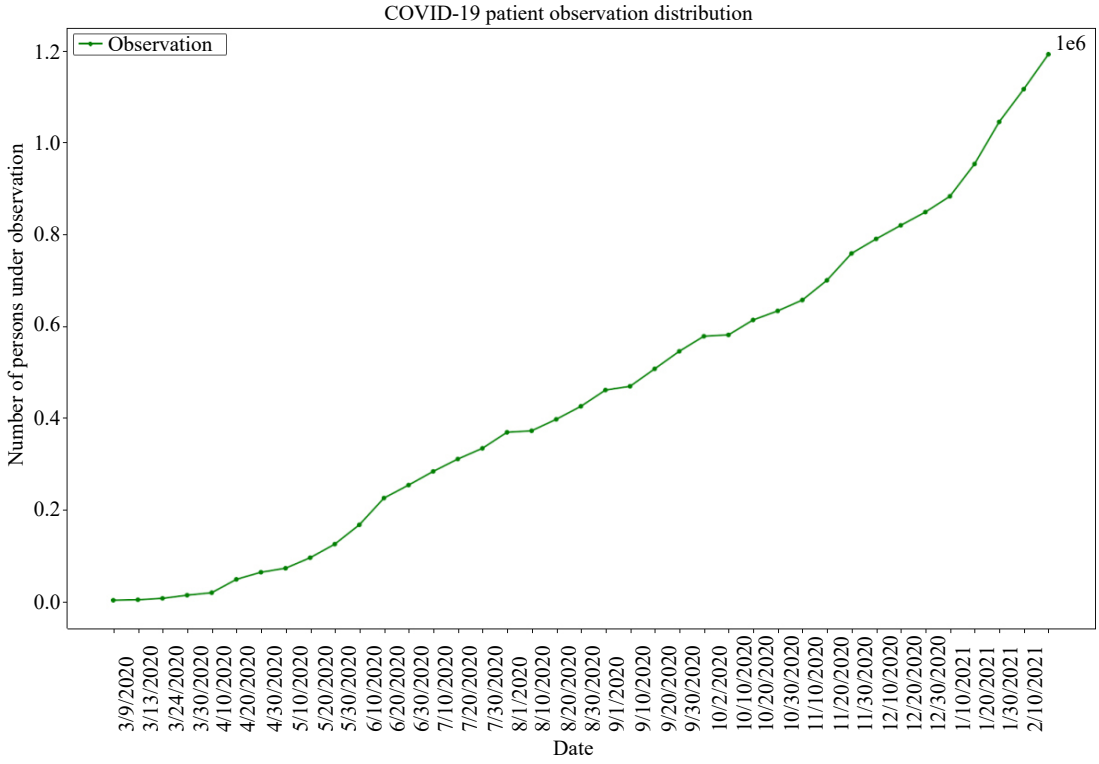
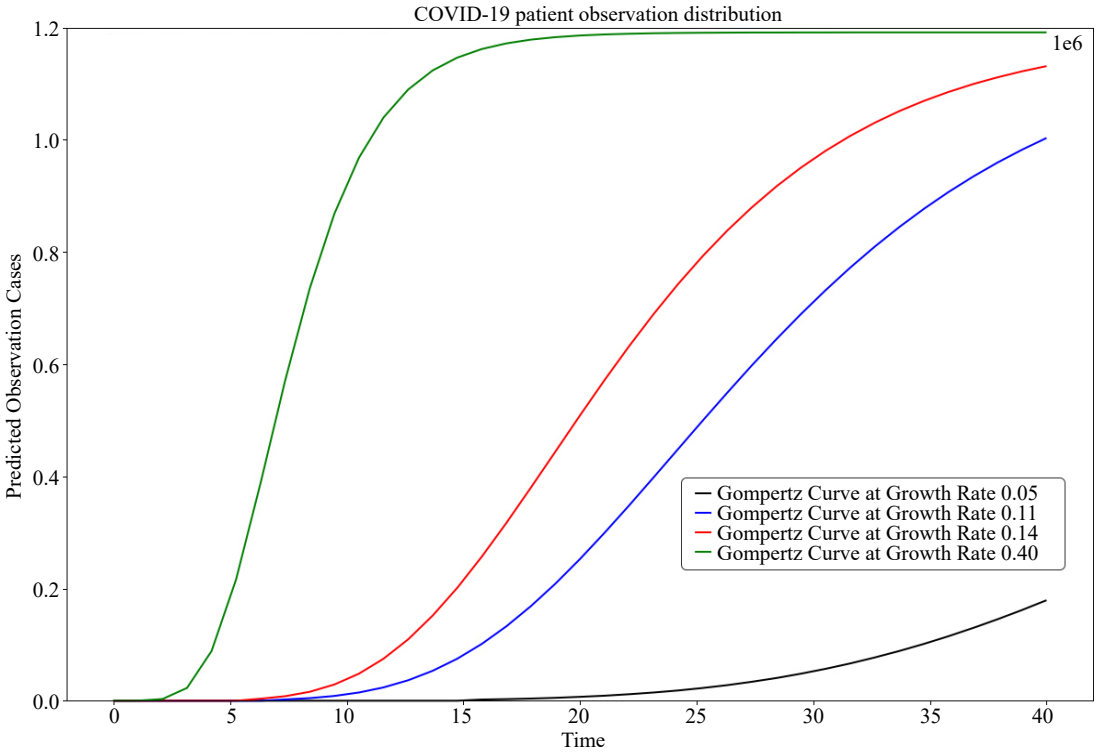


Figure 5. Original



The graphs depict the number of Persons kept under observation with respect to time. The predicted values are checked for various growth rates and the original values are calculated from the actual

Figure 6. Predicted

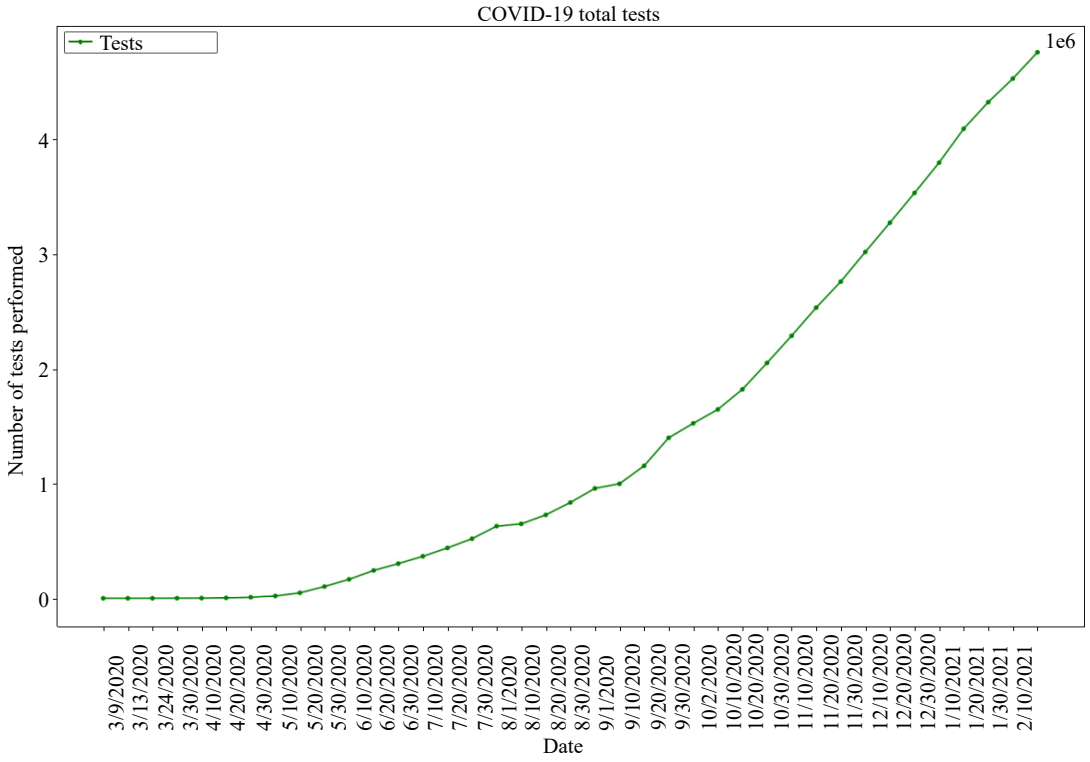
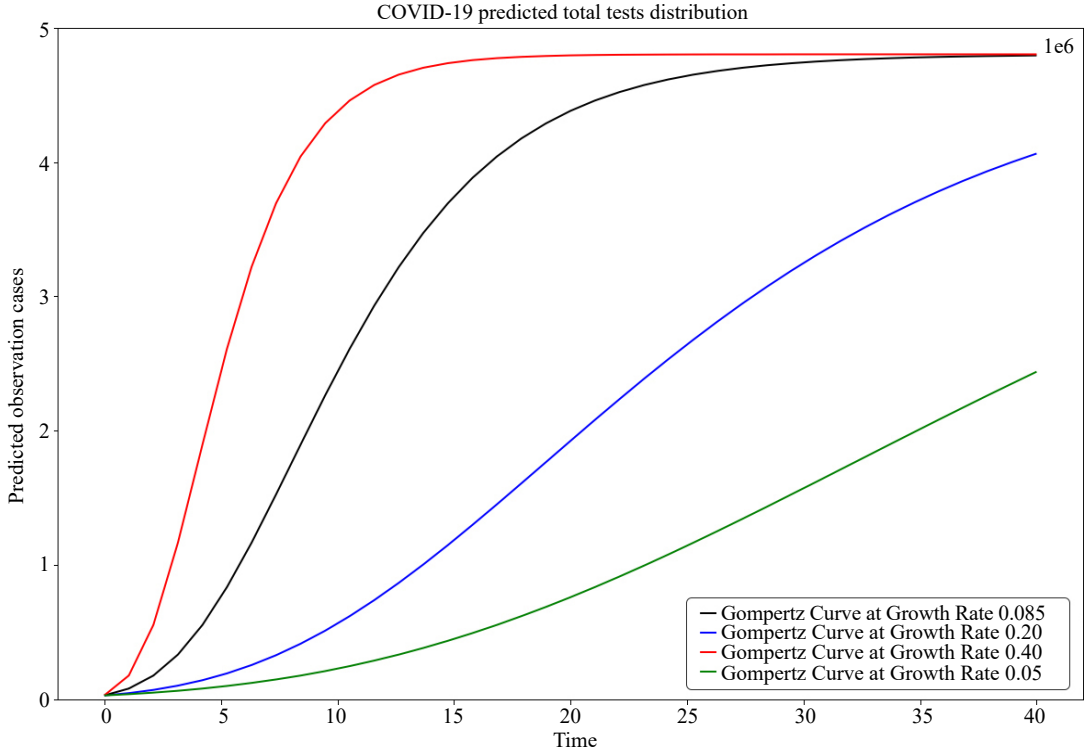


Figure 7. Original



The graphs depict the number of tests performed with respect to time. The number of patients enrolled for the tests as calculated by the actual data as well as the predicted data

Figure 8. Predicted

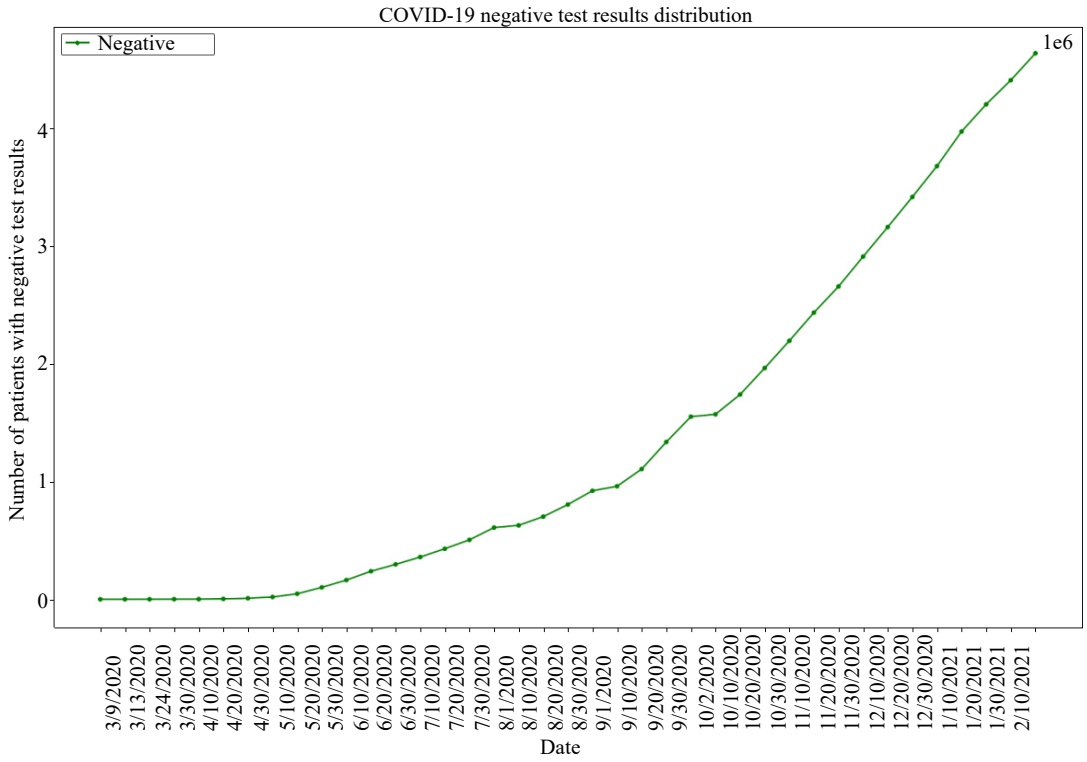
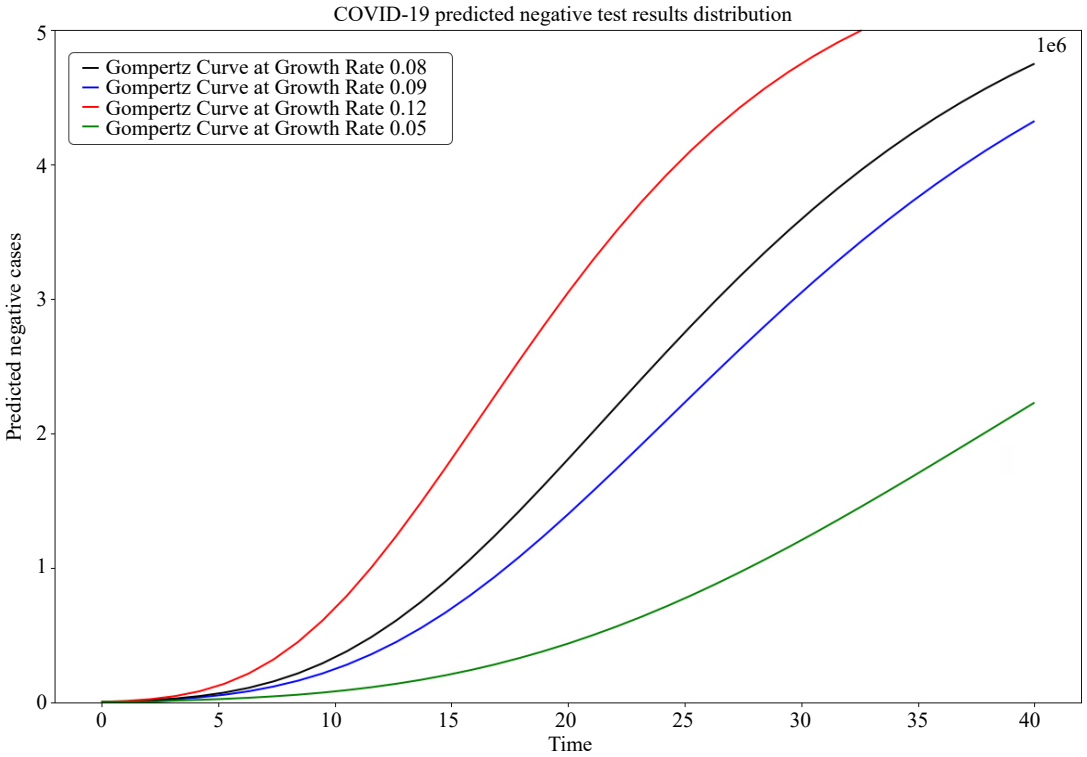


Figure 9. Original



The graphs depict the number of samples tested negative with respect to time. The graphs show the number of negative cases. The patients who are not tested positive for the disease

Figure 10. Predicted

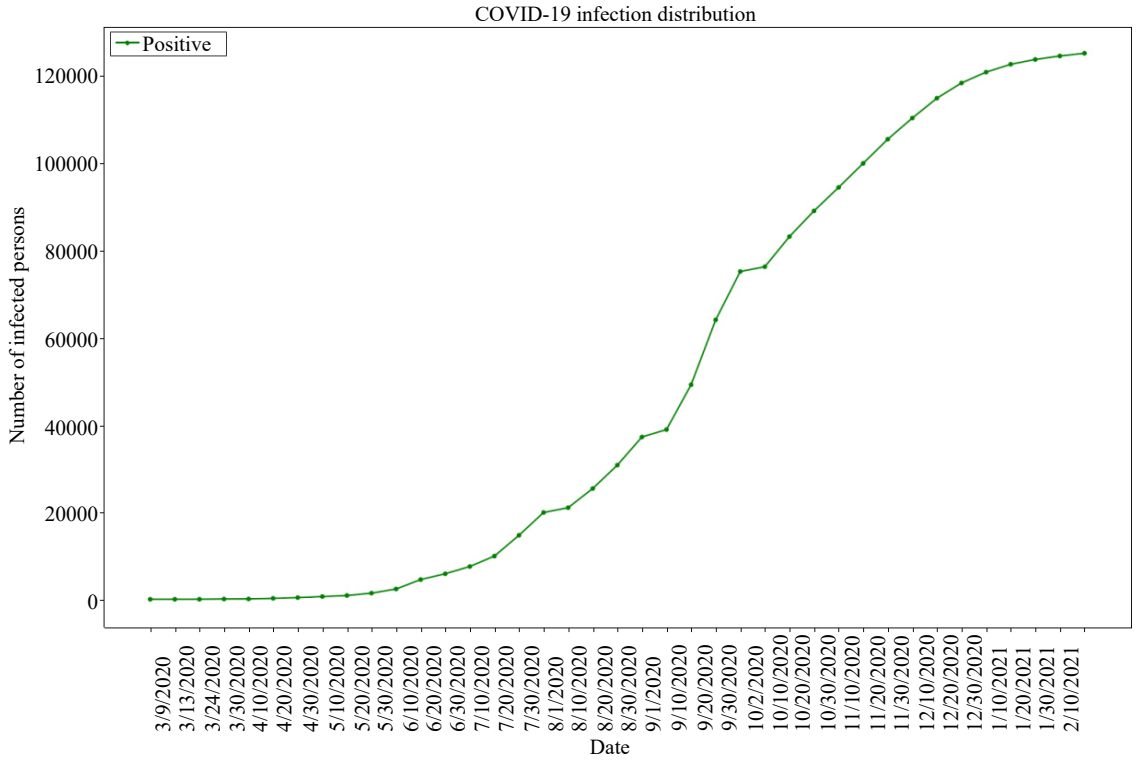
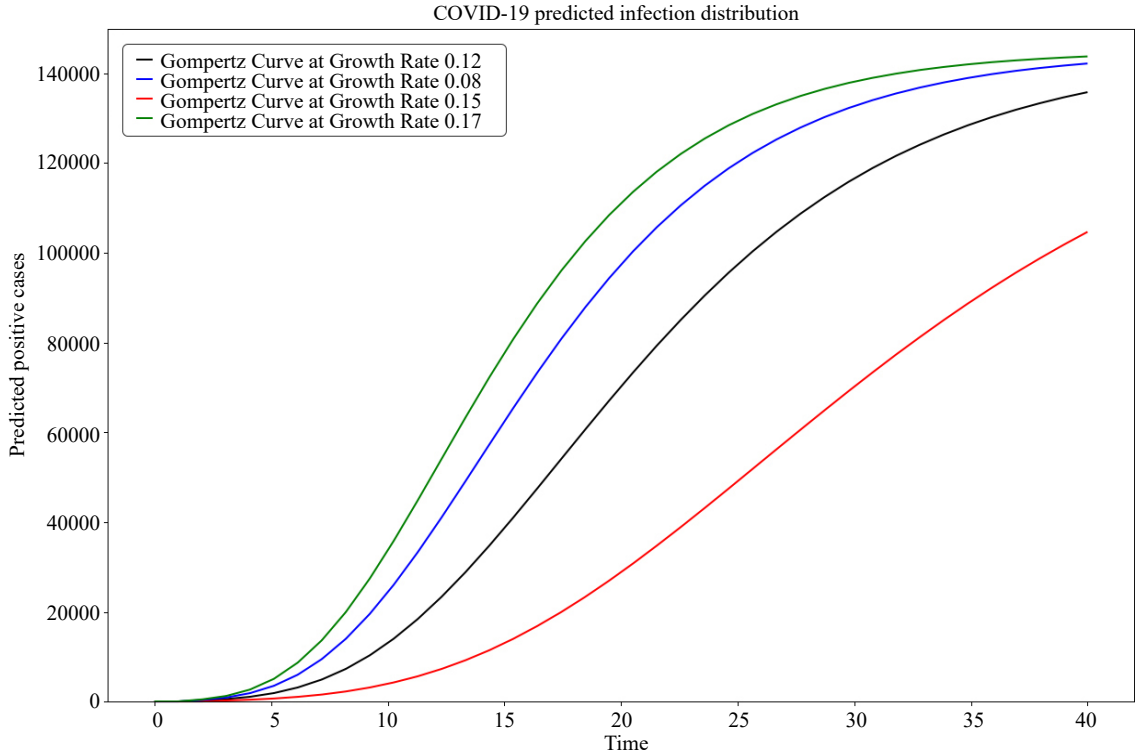


Figure 11. Original



The graphs depict the number of samples tested positive with respect to time. The graphs show the number of infected cases after proper testing. Both cases actual as well as predicted are mentioned

Figure 12. Predicted

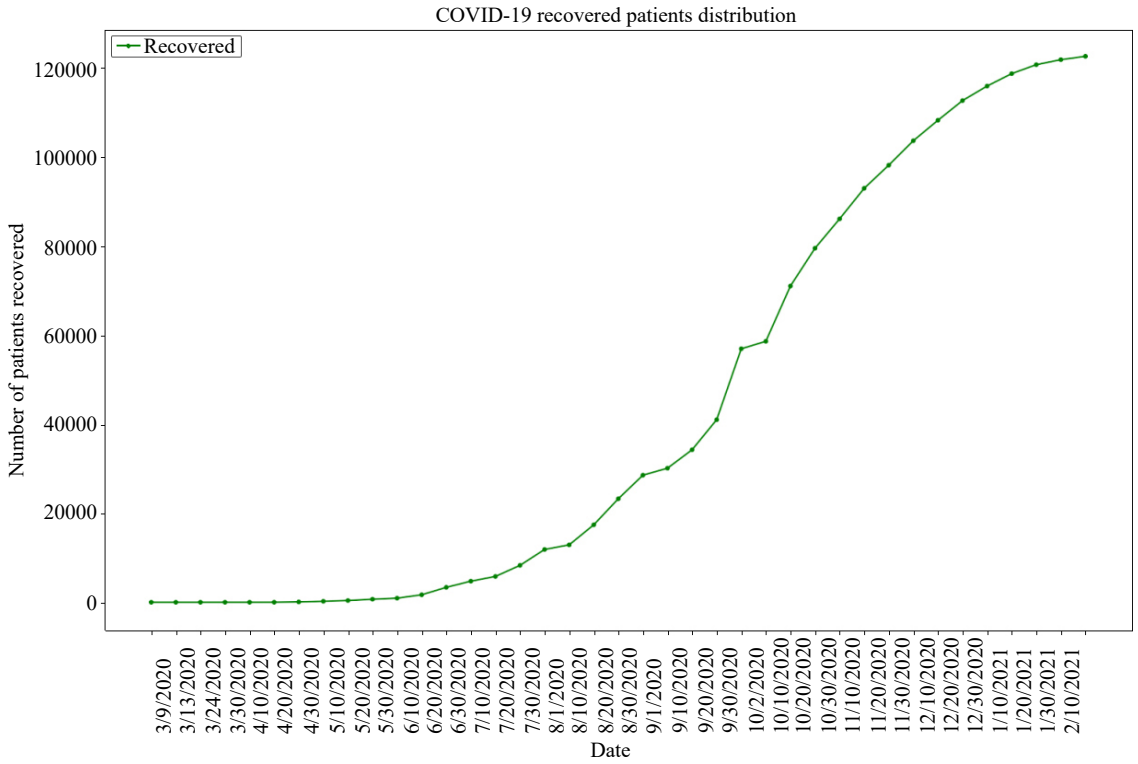
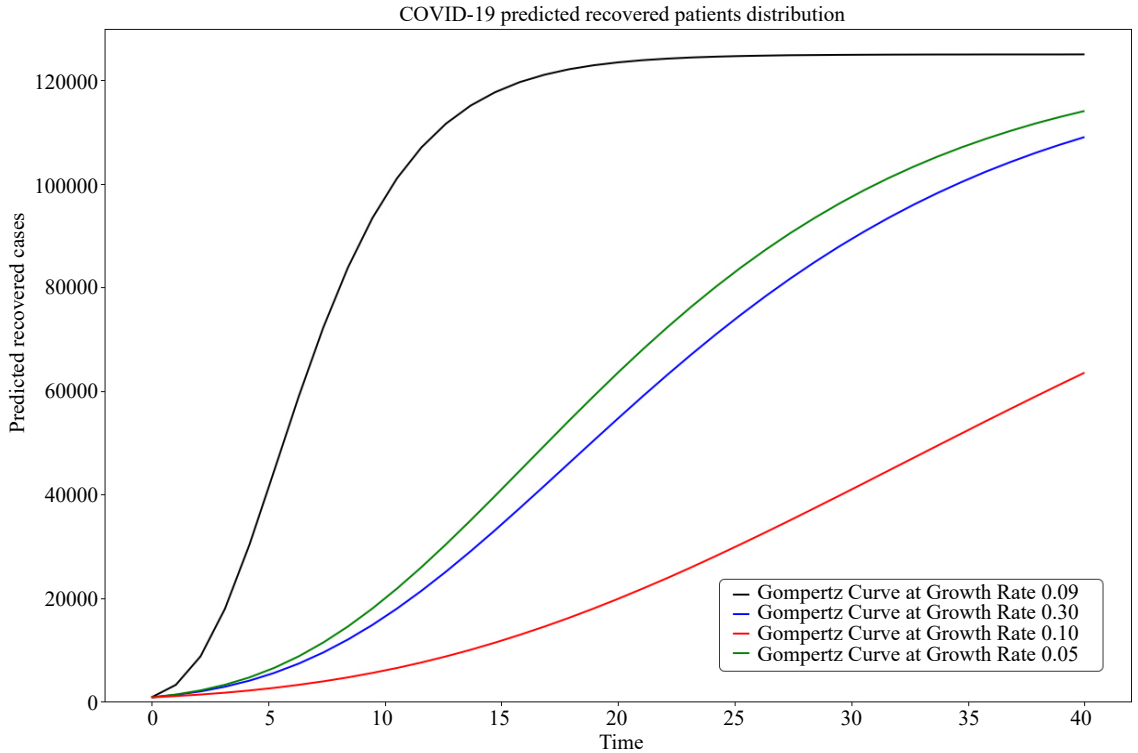


Figure 13. Original



The graphs depict the number of patients recovered with respect to time. The graphs show the number of cases recovered after being tested positive, The behaviour of the disease curve is plotted on actual data as well as predicted values

Figure 14. Predicted

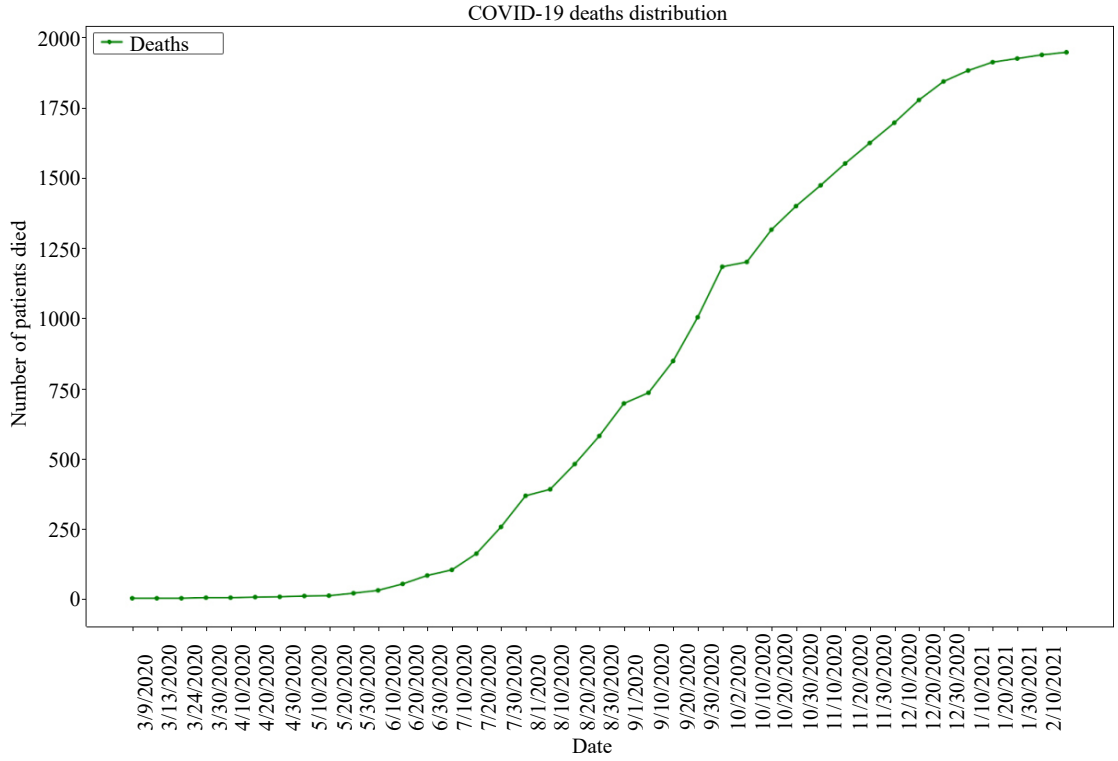
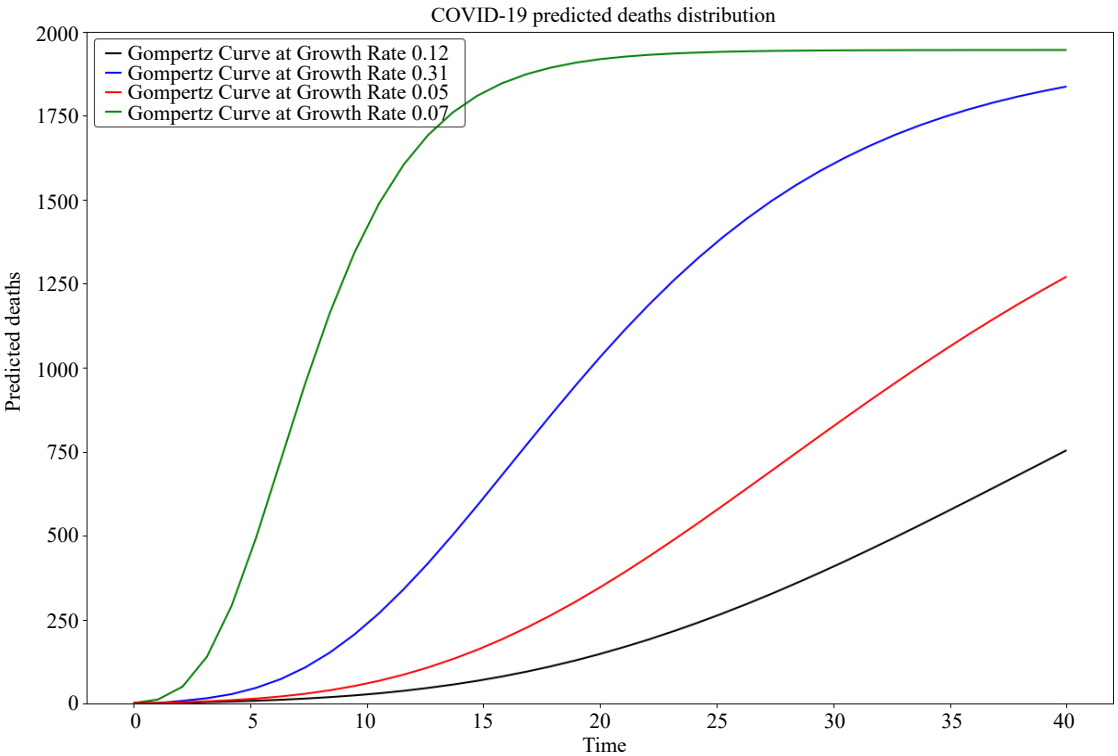


Figure 15. Original



The graphs depict the number of patients who died with respect to time. The deaths were caused due to the disease. The deaths that took place due to the disease are plotted as well as the predicted

Figure 16. Predicted

Table 3. Growth rates occurring frequently in datasets

S No.	Parameters	Growth Rate 1	Growth Rate 2	Growth Rate 3	Growth Rate 4
1	Observation	0.05	0.11	0.14	0.40
2	Tests	0.085	0.20	0.40	0.05
3	Negative	0.08	0.09	0.12	0.05
4	Positive	0.12	0.08	0.15	0.17
5	Recovered	0.09	0.30	0.10	0.05
6	Deaths	0.12	0.31	0.05	0.07

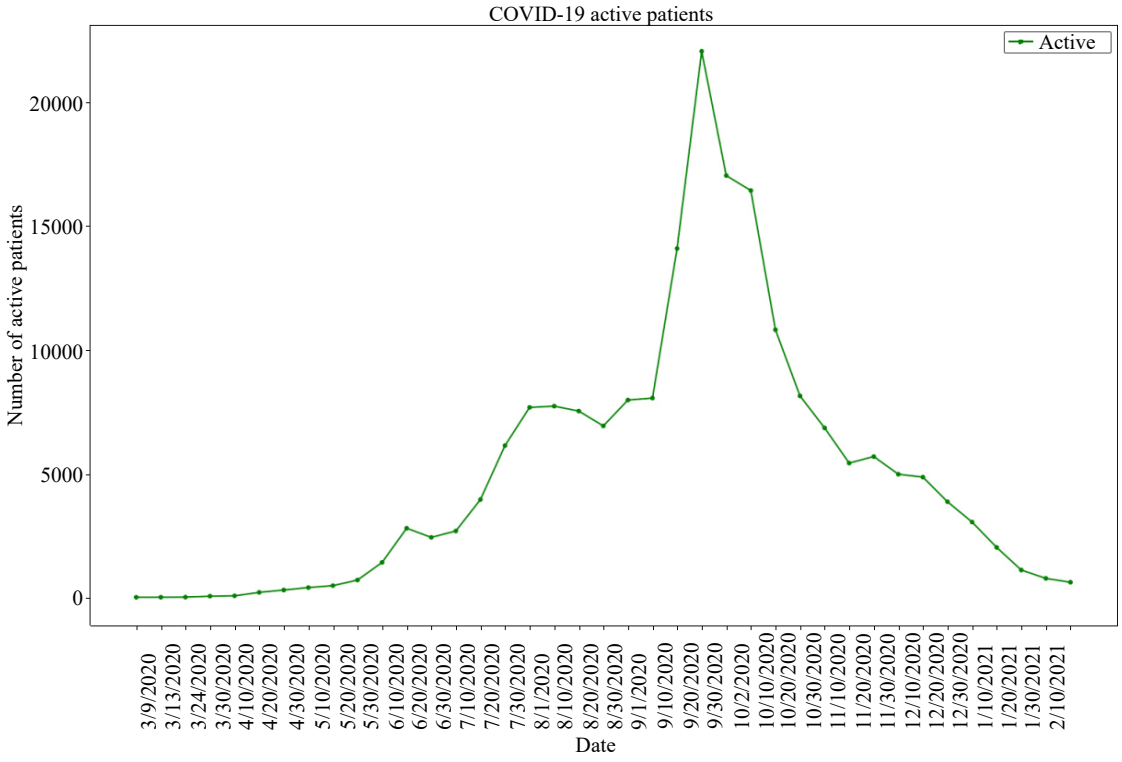


Figure 17. Original

People who are kept under observation or susceptible represent a subgroup of the population who are believed to be infected or have some symptoms of the disease. Figure 5 represents COVID-19 patients kept under observation or susceptible plotted with respect to time or Date. The graph shows a gradual increase in the number of persons under observation. Figure 6 represents the prediction-based behavior of the susceptible with respect to time. From the analysis of datasets, we observe that there are different growth rates. A few that occur mostly are chosen for the prediction are listed in Table 3.

Figure 6 consists of various curves with variable growth rates. The curve with a 0.40 growth rate is the most accurate of all. These four variable growth rates occur frequently in the dataset.

The tests are performed on a group of people kept under observation. Figure 7 shows the number of people undergoing testing for COVID-19. Figure 8 consists of four curves with variable growth rates. The curve with 0.085 growth is the closest to the original.

The outcome of the tests is either positive or negative. The samples that test negative are not infected and the samples that test positive are infected. Figure 9 and Figure 11 represent patients who tested negative and positive for COVID-19. The predicted negative and positive patients are represented in Figure 10 and Figure 12. The predicted outcomes are best represented by 0.09 and 0.15 growth rates.

The efforts to limit the pandemic by the authorities worked in reducing deaths and increasing recoveries. The curves with growth rates of 0.09 and 0.12 are the best-predicted outcomes from Figure 13 and Figure 14. Due to strict lockdown and preventive measures, the number of active cases reducing day by day can be seen in Figure 17.

Figure 18 shows the total number of positive cases generated from the actual data, the Predicted model, and SIR model; Figure 19 shows the total number of negative cases generated from the actual data, the Predicted model, and SIR model; Figure 20 shows the total number of recovered cases generated from the actual data, the Predicted model, and SIR model; Figure 21 shows the total number of deaths generated from the actual data, the Predicted model, and SIR model.

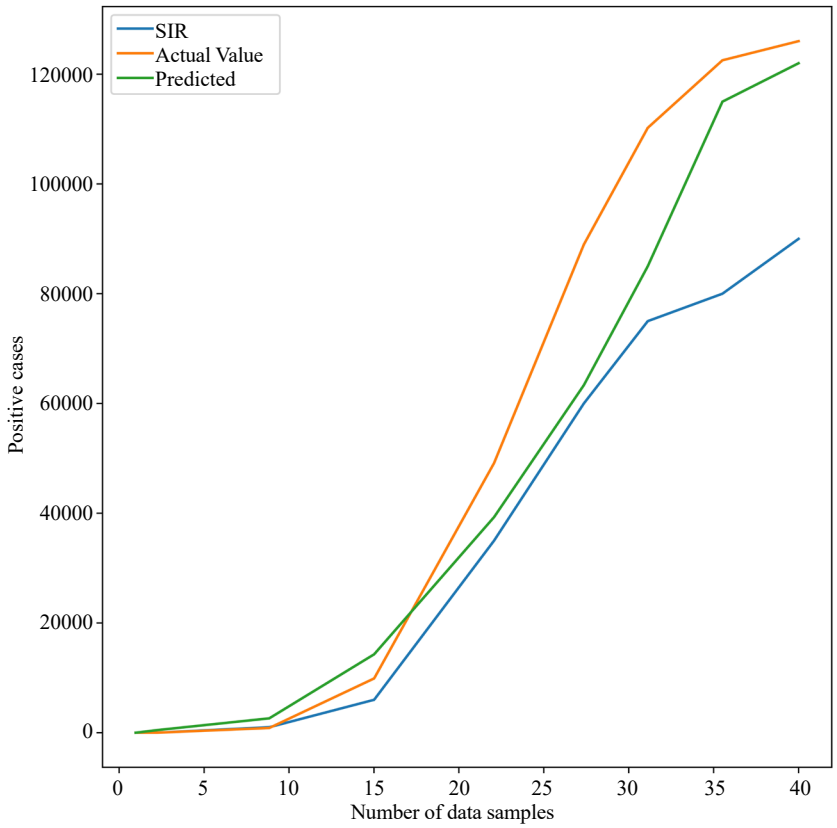


Figure 18. COVID19 positive cases

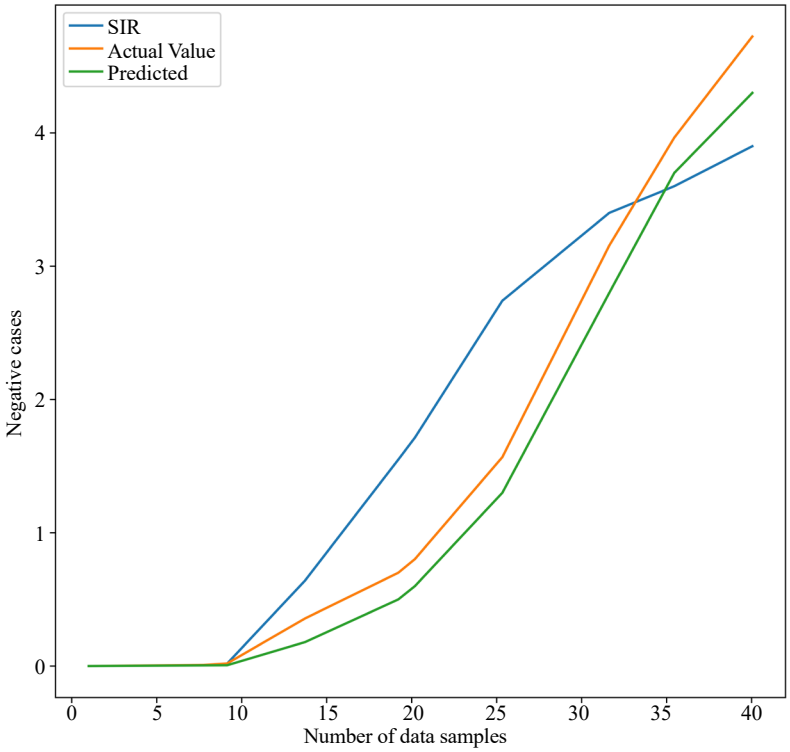


Figure 19. COVID19 negative cases

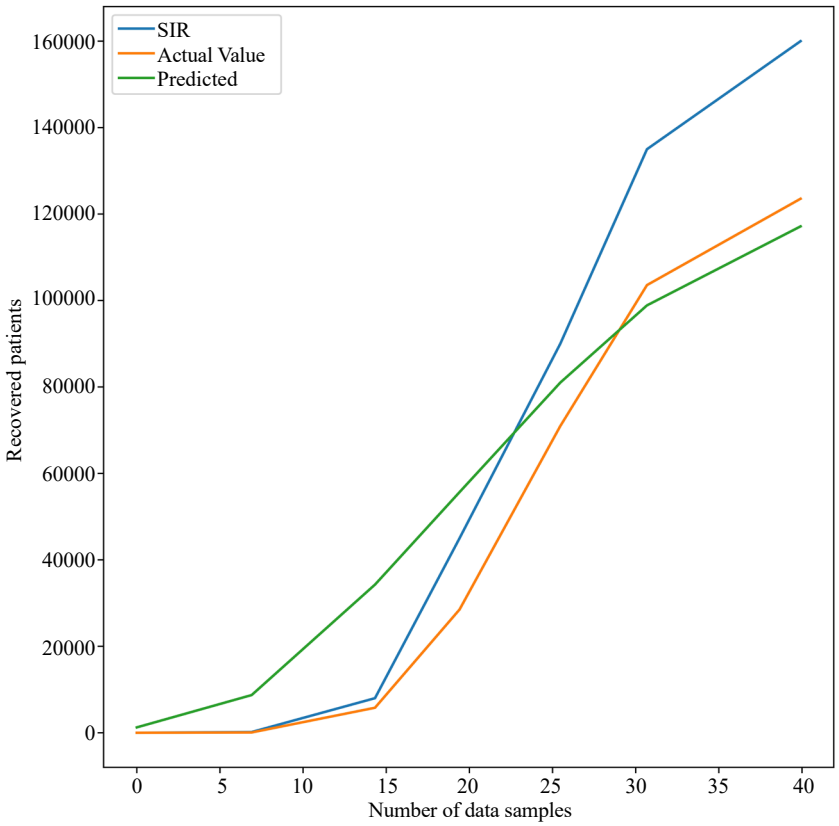


Figure 20. COVID19 recovered cases

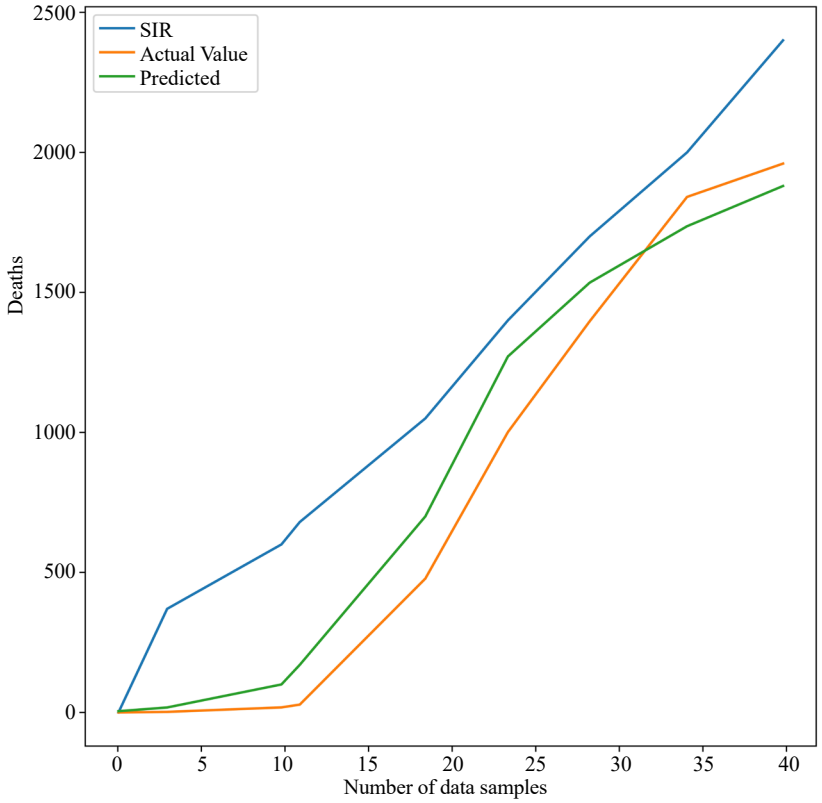


Figure 21. COVID19 death cases

9. Conclusion

In this paper, we proposed a prediction-based approach to find the epidemiological behavior of the coronavirus disease. The main aspects of the proposed approach are to find the epidemiological behavior of the pandemic through analysis and then visualization using graphs and plots. This way we can predict whether a sample is positive or negative based on the symptoms present in the patient. We assessed the proposed approach using COVID-19 datasets in Jammu and Kashmir. We integrated data from the Ministry of Health and Family Welfare, the Government of India, the Department of Information and Public Relations, and the Government of Jammu and Kashmir, and data crawled from various sources. The results predicted from the data sources were validated through the epidemiological model.

The data collected from the sources is sometimes biased due to human error. In case of any such pandemic or other conditions data must be digitized properly and with the notion to contribute to society in general and research in particular.

The work is focused on the prediction and analysis of COVID-19 disease with the help of symptoms present in the patient. The work has already helped the Government of Jammu and Kashmir during COVID-19. We encountered a lot of symptoms but our model filtered the symptoms and limited them to few. The results generated from the predicted model were validated with the actual data curve and the dataset values were also checked for the SIR model. Our model has performed better in terms of prediction. The model can be deployed in all such practical situations, where symptom-based diseases can occur in society. This work has the potential to frame public policy and decision-making at the administrative level for the benefit of society. There is a future direction for this work in terms of environmental variables. It can be a collaborative work with social science researchers. The impact of social science on developmental growth can have a huge impact on the growth of such diseases.

Availability of data and material

The data and material used in our study are properly cited and included in the bibliography section. If the readers find difficulty in getting the required links they can contact the corresponding author through the proper channel.

Funding

There does not exist any funding source to mention.

Authors contribution

Both authors have contributed equally.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Castorina P, Iorio A, Lanteri D. Data analysis on coronavirus spreading by macroscopic growth laws. *International Journal of Modern Physics C*. 2020; 31(7): 2050103.

[2] Manoj M, Kumar MS, Valsaraj K, Sivan C, Vijayan SK. Potential link between compromised air quality and transmission of the novel corona virus (SARS-CoV-2) in affected areas. *Environmental Research*. 2020; 190: 110001.

[3] Sahoo PK, Powell MA, Mittal S, Garg V. Is the transmission of novel coronavirus disease (COVID-19) weather dependent? *Journal of the Air & Waste Management Association*. 2020; 70(11): 1061-1064.

[4] Wang J, Tang K, Feng K, Lin X, Lv W, Chen K, et al. High temperature and high humidity reduce the transmission of COVID-19. *Available at SSRN*. 2020; 57: 1-13.

[5] La Rosa G, Bonadonna L, Lucentini L, Kenmoe S, Suffredini E. Coronavirus in water environments: Occurrence, persistence and concentration methods-a scoping review. *Water Research*. 2020; 179: 115899.

[6] Hamidi S, Sabouri S, Ewing R. Does density aggravate the COVID-19 pandemic? *Journal of the American Planning Association*. 2020; 86(4): 495-509.

[7] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. *New England Journal of Medicine*. 2020; 382(13): 1199.

[8] Kroumpouzos G, Gupta M, Jafferany M, Lotti T, Sadoughifar R, Sitkowska Z, et al. COVID-19: A relationship to climate and environmental conditions? *Dermatologic Therapy*. 2020; 33(4): e13399.

[9] Bu J, Peng DD, Xiao H, Yue Q, Han Y, Lin Y, et al. *Analysis of meteorological conditions and prediction of epidemic trend of 2019-nCoV infection in 2020*. MedRxiv. 2020. Available from: https://doi.org/10.1101/2020.02.13.20022715.

[10] Wu Y, Jing W, Liu J, Ma Q, Yuan J, Wang Y, et al. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Science of the Total Environment*. 2020; 729: 139051.

[11] Shao N, Zhong M, Yan Y, Pan H, Cheng J, Chen W. Dynamic models for coronavirus disease 2019 and data analysis. *Mathematical Methods in the Applied Sciences*. 2020; 43(7): 4943-4949.

[12] Yin R, Feng W, Wang T, Chen G, Wu T, Chen D, et al. Concomitant neurological symptoms observed in a patient diagnosed with coronavirus disease 2019. *Journal of Medical Virology*. 2020; 92(10): 1782.

[13] Volpert V, Banerjee M, Petrovskii S. On a quarantine model of coronavirus infection and data analysis. *Mathematical Modelling of Natural Phenomena*. 2020; 15: 24.

[14] Phucharoen C, Sangkaew N, Stosic K. The characteristics of COVID-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data. *EClinicalMedicine*. 2020; 27: 100543.

[15] Garcia LP, Goncalves AV, de Andrade MP, Pedebos LA, Vidor AC, Zaina R, et al. Estimating underdiagnosis of COVID-19 with nowcasting and machine learning. *Revista Brasileira de Epidemiologia*. 2021; 24: 1-13.
[16] Flesia L, Monaro M, Mazza C, Fietta V, Colicino E, Segatto B, et al. Predicting perceived stress related to the COVID-19 outbreak through stable psychological traits and machine learning models. *Journal of Clinical Medicine*. 2020; 9(10): 3350.
[17] Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. *Jama*. 2020; 323(14): 1406-1407.
[18] Di Castelnuovo A, Bonaccio M, Costanzo S, Gialluisi A, Antinori A, Berselli N, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre italian corist study. *Nutrition, Metabolism and Cardiovascular Diseases*. 2020; 30(11): 1899-1913.
[19] Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*. 2020; 139: 110059.
[20] Al-Rousan N, Al-Najjar H. Data analysis of coronavirus COVID-19 epidemic in south korea based on recovered and death cases. *Journal of Medical Virology*. 2020; 92(9): 1603-1608.
[21] Dey SK, Rahman MM, Siddiqi UR, Howlader A. Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. *Journal of Medical Virology*. 2020; 92(6): 632-638.
[22] Hamzah FB, Lau C, Nazri H, Ligot D, Lee G, Tan C, et al. Coronatracker: worldwide COVID-19 outbreak data analysis and prediction. *Bull World Health Organ*. 2020; 1(32): 1-32.
[23] Amar LA, Taha AA, Mohamed MY. Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt. *Infectious Disease Modelling*. 2020; 5: 622-634.
[24] Manevski D, Gorenjec NR, Kejžar N. Modeling COVID-19 pandemic using bayesian analysis with application to slovene data. *Mathematical Biosciences*. 2020; 329: 108466.
[25] Bashir MF, Ma B, Komal B, Bashir MA, Tan D, Bashir M. Correlation between climate indicators and COVID-19 pandemic in NEW YORK, USA. *Science of the Total Environment*. 2020; 728: 138835.
[26] Chan KH, Peiris JM, Lam SY, Poon LLM, Yuen KY, Seto WH. The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Advances in Virology*. 2011; 2011: 1-8.
[27] Sahoo PK, Chauhan AK, Mangla S, Pathak AK, Garg VK. COVID-19 pandemic and its relationship with environment impacts and factors: An outlook from punjab and chandigarh, India. *Environmental Forensics*. 2021; 22(1-2): 143-154.
[28] Manoj MG, Kumar MS, Valsaraj KT, Sivan C, Vijayan SK. Potential link between compromised air quality and transmission of the novel corona virus (SARS-CoV-2) in affected areas. *Environmental Research*. 2020; 190: 110001.
[29] Shi P, Dong Y, Yan H, Zhao C, Li X, Liu W, et al. Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Science of the Total Environment*. 2020; 728: 138890.
[30] Xie J, Zhu Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Science of the Total Environment*. 2020; 724: 138201.
[31] Jamil T, Alam I, Gojobori T, Duarte CM. No evidence for temperature-dependence of the COVID-19 epidemic. *Frontiers in Public Health*. 2020; 8: 436.
[32] Mourtzoukou E, Falagas ME. Exposure to cold and respiratory tract infections. *The International Journal of Tuberculosis and Lung Disease*. 2007; 11(9): 938-943.
[33] Wani MA, Wani DM, Naik S, Mayer IA. Geographical vulnerability to respiratory infections using GIS technique-micro analysis study in the Himalayan region-India. *GeoJournal*. 2020; 87: 1193-1215.
[34] Koul PA, Mir MA, Bali NK, Chawla-Sarkar M, Sarkar M, Kaushik S, et al. Pandemic and seasonal influenza viruses among patients with acute respiratory illness in Kashmir (India). *Influenza and Other Respiratory Viruses*. 2011; 5(6): e521-e552.
[35] Dataset Sources. Available from: https://www.jkinfonews.com/index.aspx,http://new.jkdirinf.in/,https://twitter.com/diprjk?ref src=twsrc%5Egoogle% 7Ctwcamp%5Eserp%7Ctwgr%5Eauthor [Accessed 11 February 2021].
[36] Dataset Sources. Available from: https://www.mohfw.gov.in/ [Accessed 11 February 2021].
[37] Dataset Sources. Available from: https://www.worldometers.info/coronavirus/country/india/ [Accessed 11 February 2021].
[38] Dataset Sources. Available from: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge [Accessed 11 February 2021].