REPORT

on

REAL-TIME FORECASTING OF AN EPIDEMIC USING DISCRETE TIME STOCHASTIC MODEL: A CASE STUDY OF PANDEMIC COVID-19

May 30, 2020

Shivanagouda Biradar
Under the guidance of
Shubhendu Bhashin
IIT Delhi
Department of Electrical Engineering

Contents

1 Abstract

Corona virus disease 2019 (COVID-2019) has been recognized as a pandemic, and several studies are being conducted using various mathematical models to predict the probable evolution of this epidemic. These mathematical models based on various factors, and analysis are subject to potential errors. Here, we show that a simple probabilistic model could be useful to predict the spread of COVID-2019.

This work recognises the stochastic nature of the epidemic dynamics and uses the concepts of well established stochastic models like Reed-Frost model and Greenwood Model, to predict the epidemiological trend of the prevalence and incidence of COVID-2019.

Two algorithms have been used to predict incidences of COVID-19. Both Algorithm-1 and Algorithm-2, uses the Greenwood stochastic model, a particular case of generalised chain binomial model.

In Algorithm-1, branching process approximation has been used, by assuming the exponential growth of cases within each reporting interval. The epidemic curve is predicted using only two parameters.

Algorithm-2 is based on a premise that a person in susceptible population can either be infected or escapes the infection with certain probability when he comes in adequate contact with an infective. This process creates a chain of infection in population. The Algorithm-2 models this chain of infection as a binomial process, and estimates the probabilities, which can be further used to predict incidence of COVID-19.

2 Basics

The population comprises of two set of individuals. Those who are already infected are called as infectives and those capable of receiving the infection are called susceptible. The models assumes that all the individuals in population have equal probability of being susceptible and spreading the disease. Before moving further into discussion, let us define some important terms:

- **Latent or Incubation time period:** The period of time during which development of disease is purely internal to the infected person.
- **Infectious period:** Time period during which infective can transmit the disease to member of susceptible population.

- λ_t**:** Probability of susceptible escaping the infection in time interval $[t-1, t]$.
- **Generation time:** The time interval between infection of primary case and infection of secondary case caused by primary case.
- **Reproduction number:** The basic reproduction number (R_0) is used to measure the transmission potential of a disease. It is the average number of secondary infections produced by a typical case of an infection in a population where everyone is susceptible.

Now, let N be the size of population. The disease start with certain number of infected individuals I_0 at $t=0$.The remaining $N-I_0$ members of populations are susceptible. Let S_t denotes the number of susceptible, I_t denote the number of infectives and D_t denotes the number of recovered plus dead individuals at time t in the population (in this work t represents day t, not time t). The recursive relationship

$$S_t = C_{t+1} + S_{t+1} \tag{1}$$

$$I_{t+1} = I_t + C_{t+1} - D_{t+1} \tag{2}$$

for $t = 0, 1, \ldots$ and $N = S_0 + I_0$ holds.

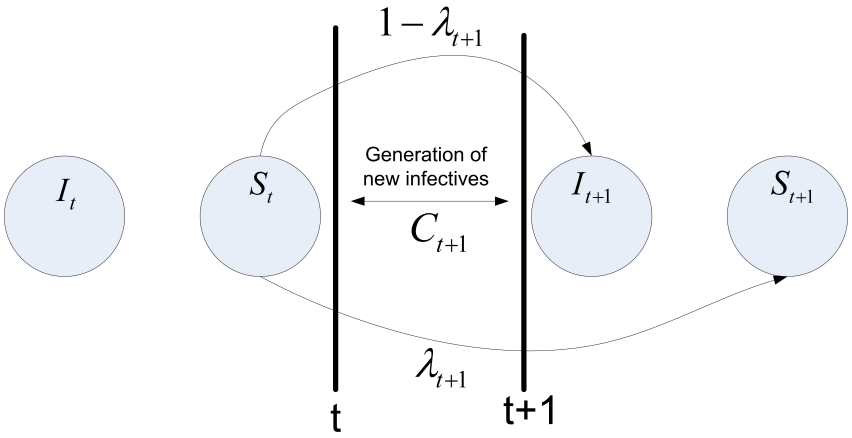

Figure 1

If the susceptibles S_t, can be regarded as a binomial random variable, S_t can either become infected with probability $1-\lambda_{t+1}$ and converted to infective I_{t+1} or remains susceptible and is converted to S_{t+1}, as represented in Figure 1. Here, the probability of generation of new infectives can be thought to be based on two following assumptions:

- **Assumption-1**: The probability of generation of new infectives is dependent on the number of infectives in the population at time t.
- **Assumption-2**: The probability of generation of new infectives is independent of the number of infectives in the population at time t.

Based on these assumption, two models were developed. These models are well established and used in epidemiological community with great confidence.

2.1 Reed-Frost & Greenwood Model

Reed and Wade H. Frost in lectures at Johns Hopkins University developed the first chain binomial epidemic model using assumption-1, i.e. the probability of generation of new infective during $[t, t+1]$ depends on the number of infectives present in the population. Therefore, the probability of having C_{t+1} infected individual from a population of S_t individuals conditioned on infectives at t, is given as:

$$P(C_{t+1} = c_{t+1} \,|\, I_t = i_t, S_t = s_t) = \binom{s_t}{c_{t+1}} \left(1 - \lambda_{t+1}^{c_t}\right)^{c_{t+1}} \left(\lambda_{t+1}^{c_t}\right)^{s_{t+1}} \tag{3}$$

$$= \frac{s_t!}{s_{t+1}! c_{t+1}!} \left(1 - \lambda_{t+1}^{c_t}\right)^{c_{t+1}} \left(\lambda_{t+1}^{c_t}\right)^{s_{t+1}}. \tag{4}$$

Here, $1 - \lambda^{c_t}$ is the probability of susceptible getting infected by at least one infectives. The Greenwood model is a based on assumption-2, then the corresponding probability of having C_{t+1} individuals is expressed as:

$$P(C_{t+1} = c_{t+1} \,|\, I_t = i_t, S_t = s_t) = \binom{s_t}{c_{t+1}} (1 - \lambda_{t+1})^{c_{t+1}} (\lambda_{t+1})^{s_{t+1}} \tag{5}$$

$$= \frac{s_t!}{s_{t+1}! c_{t+1}!} (1 - \lambda_{t+1})^{c_{t+1}} (\lambda_{t+1})^{s_{t+1}}. \tag{6}$$

Notice that the Greenwood model is slightly simpler than Reed-Frost model due the assumption.

Two models were combined to form a Generalised chain binomial model, which was developed by Neils Becker in 1981. Similar to the previous models, we express the probability of having C_{t+1} individuals as:

$$P(S_{t+1}=x, C_{t+1}=s-x \mid S_t=s, I_t=i_t)=\binom{s}{x}(1-q_t)^{s-x}q_t^x, \forall x=0,1,\ldots,s. \tag{7}$$

Where, if $q_t=\lambda_t^{c_t}$, then we get Reed-Frost model and if $q_t=\lambda_t$, we get Greenwood model.

These models in general are advantageous since they help in following the course of epidemic through population by following epidemic chain. For sake of better understanding let us consider a chain of infectives represented by $1-2-1-0$, each number represents the number of infectives in each generation. The conditional probability of generating this chain is given by:

$$P(1-2-1)=P(S_1=2 \mid S_0=4, I_0=1)\,P(S_2=1 \mid S_1=2, I_1=2)\,P(S_3=0 \mid S_2=1, I_2=1)$$

$$=\binom{4}{2}q_1^2p_1^2\binom{2}{1}q_2^1p_2^1\binom{1}{0}q_3^0p_3^0 \tag{8}$$

$$=12q_1^2p_1^2q_2p_2$$

The above equation can be transformed to Reed-Frost model by using transformation $q_i=\lambda_i^{c_t}$, we get $12\left(\lambda_1^2\right)^2\left(1-\lambda_1^2\right)^2\lambda_2(1-\lambda_2)$. Similarly, the Greenwood model is obtained by using transformations $q_i=\lambda_i$, we get $12(\lambda_1)^2(1-\lambda_1)^2\lambda_2(1-\lambda_2)$.

In the general chain binomial model, it is possible to write the probability of any chain $I_0-C_1-C_2-\ldots-C_r$ of infectives for times $t=0,1,2\ldots$ as:

$$P(I_0-C_1-\cdots-C_r)=\frac{S_0!}{C_1!C_2!\ldots C_r!}\prod_{t=0}^{r}q_{i_t}^{s_{t+1}}p_{i_t}^{C_{t+1}} \tag{9}$$

Where, $S_t=S_0-\sum_{i=0}^{t}C_t$.

It is possible to enumerate and calculate all possible chains and their associated probabilities for small values for N.

The Reed- Frost and Greenwood models require estimates of the basic parameter q of the model. This is accomplished by using maximum likelihood methods.

Some of the questions worth asking at this point of time is,

- What if the susceptible population is very large?
- Is it possible to build model for such a large chain?

- Can we approximate the considered binomial model such that large chain can be taken into account?.

3 Algorithm-1

3.1 Model Building Process:

Given S_t and C_t, the chain binomial model predicts S_{t+1} and C_{t+1} iteratively using binomial distributions. The deterministic expressions can be written as:

$$E[C_{t+1}] = S_t(1-\lambda_{t+1}) \tag{10}$$

$$E[S_{t+1}] = S_t - E[C_{t+1}] \tag{11}$$

Equation (10), is obtained from applying expectation operator on Equation (1) and using the fact that $E[S_{t+1}] = S_t\lambda_{t+1}$ for large susceptible population. Equation (11) is also obtained from equation (1).
Here, λ_{t+1} is the probability that susceptible individuals S_t escape infection on **day** t.
The time series of new incidences of infectives can be written as:

$$C_{t+1} \sim Binomial(S_t, 1-\lambda_{t+1}). \tag{12}$$

Equation(12) can be more precisely written as (Greenwood Model):

$$P(C_{t+1} = x | S_t = s, \lambda_{t+1}) = \frac{s!}{x!(s-x)!}(1-\lambda_{t+1})^x(\lambda_{t+1})^{s-x} \tag{13}$$

The probability of obtaining C_t incidence can be conditioned on the previous incidences, given as:

$$P(C_0 \to C_1 \cdots \to C_t) = \frac{S_0!}{C_0! \cdots C_t! S_t!}\prod_{i=0}^{t-1}(1-\lambda_{i+1})^{C_{i+1}}\lambda_{i+1}^{S_t} \tag{14}$$

Where $S_t = S_0 - \sum_{i=0}^{t} C_t$.
Despite the usefulness of Greenwood model, it is not applied to large susceptible population. Instead, a special case of binomial model is used i.e. Possion distribution model. Poisson distribution describes the probability of occurrence of an event in a time period, which is described by its parameter. Therefore, redefining the above binomial model as:

$$P(C_{t+1} = x | C_t, \ldots, C_0, S_0) = \frac{(E[C_{t+1}])^x e^{-E[C_{t+1}]}}{x!}. \tag{15}$$

We now define $E[C_{t+1}]$, using an **approximate branching process**, which tries to project the incidence on t^{th} day to $t+1^{th}$ day, using exponential growth between interval $[t, t+1]$. It is very important to take into account the fact that many generation of cases exist between the reporting interval $[t, t+1]$. For instance, a contact tracing of H1N1- 2009 in the Netherlands estimated the mean generation time as $T_g = 2.7$ days, implying that weekly data can include more than two generations of influenza cases. Therefore, we impose an exponentially increasing argument to the dynamics within each reporting interval, since no information regarding the dynamics within interval is available.

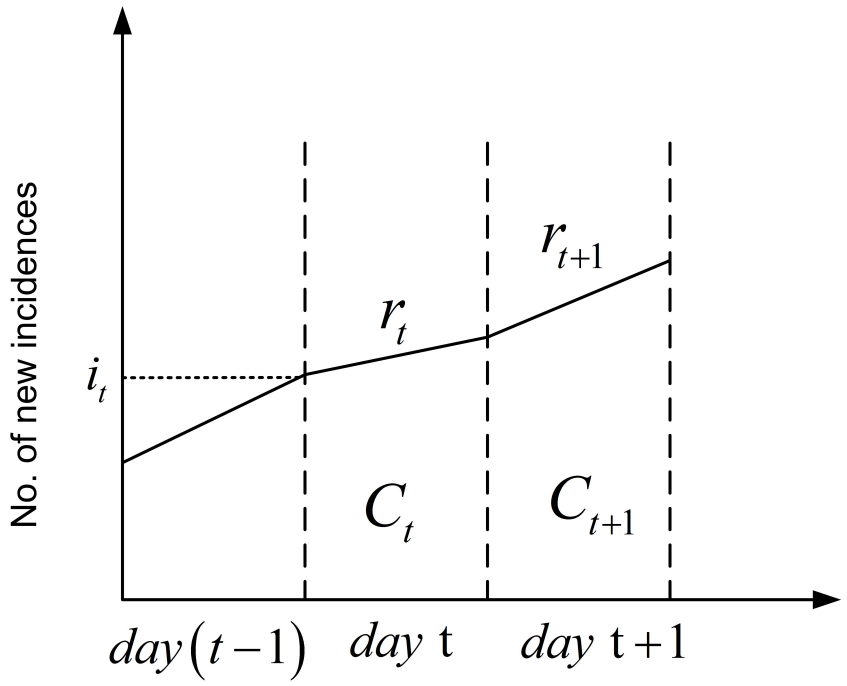

Figure 2

Exponential growth between days with a growth rate r_t is assumed. The area under the epidemic curve (Figure 2) in day t corresponds to the reported daily incidence C_k.

$$E[C_t] = \int_0^{\Delta t} i_t e^{r_t \tau} d\tau \tag{16}$$

$$= \frac{i_t}{r_t}\left[e^{r_t \Delta t} - 1\right] \tag{17}$$

and

$$E[C_{t+1}] = \int_0^{\Delta t} i_t e^{r_t \Delta t} e^{r_{t+1}\tau} d\tau \tag{18}$$

$$= \frac{i_t e^{r_t \Delta t}}{r_{t+1}} \left[e^{r_{t+1}\Delta t} - 1 \right] \tag{19}$$

from equation (17) and equation (19) we get a recursive relation as given below:

$$E[C_{t+1}] = \frac{r_t e^{r_t \Delta t}}{r_{t+1}} \frac{e^{r_{t+1}\Delta t} - 1}{e^{r_t \Delta t}} E[C_t]. \tag{20}$$

We define, the instantaneous reproduction number R_t in terms of initial reproduction number R_0 as :

$$R_t = \frac{S_{t-1}}{S_0} R_0 \tag{21}$$

Moreover, assuming that the generation time is a constant T_g days, then an estimator of the instantaneous reproduction number R_t is :

$$R_t = e^{r_t T_g} \tag{22}$$

From equation (21) and equation (22), we get:

$$r_t = \frac{1}{T_g} \ln\left(\frac{S_{t-1}}{S_0} R_0 \right) = \ln\left(\frac{S_{t-1}}{S_0} R_0 \right)^{\frac{1}{T_g}}. \tag{23}$$

Using the above equation(23) in equation(20), we get:

$$E[C_{t+1}] = \frac{\left(\frac{S_{t-1}}{S_0} R_0 \right)^{\frac{\Delta t}{T_g}} \left[\left(\frac{S_t}{S_0} R_0 \right)^{\frac{\Delta t}{T_g}} - 1 \right] \ln\left(\frac{S_{t-1}}{S_0} R_0 \right)^{\frac{1}{T_g}}}{\left[\left(\frac{S_{t-1}}{S_0} R_0 \right)^{\frac{\Delta t}{T_g}} - 1 \right] \ln\left(\frac{S_t}{S_0} R_0 \right)^{\frac{1}{T_g}}} E[C_t] \tag{24}$$

$$= A_t E[C_t] \tag{25}$$

Where $S_t = S_0 - \sum_{i=0}^{t} C_i$.

Now, substitute $E[C_{t+1}] = A_t E[C_t]$ in equation(15), we get:

$$P(C_{t+1} = x | C_t, \ldots, C_0, S_0) = \frac{(A_t C_t)^x e^{-A_t C_t}}{x!} \tag{26}$$

Assuming T_g is known, the epidemic curve is governed by only two unknown parameters S_0 and R_0.

3.2 Statistical Estimation of Unknown Parameters

The estimation of unknown parameters S_0 and R_0 given daily incidence data $C_0, \dots C_t$ can be calculated using Maximum likelihood estimation. The likelihood function is defined as:

$$L(S_0, R_0 | C_0, \cdots, C_t) = \prod_{h=1}^{t} \frac{(E[C_h])^{C_h} e^{-(E[C_h])}}{C_h!} \tag{27}$$

The estimates are obtained by maximizing the negative logarithm of equation (27).

3.3 Results

Data Window (data used for prediction of parameters)	Initial reproduction number R_0	Initial susceptible population(Million) S_0	
30-Jan-2020 to 9-April-2020	1.4	1580	Figure-3
30-Jan-2020 to 9-May-2020	1.23	1350	Figure-4
30-Jan-2020 to 21-May-2020	1.2	1390	Figure-5

In Figure 3, 4 and 5, the **filled** dots represent the data window i.e. the data used for estimating the parameters. The **unfilled** dots represents the observed data. The **black dotted** line is the estimated curve.
In all figures, blue color represents 5 days ahead prediction starting from the end of data window, cyan color represents 10 days ahead prediction starting from the end of data window, yellow color represents 20 days ahead prediction starting from the end of data window, green color represents 30 days ahead prediction starting from the end of data window.

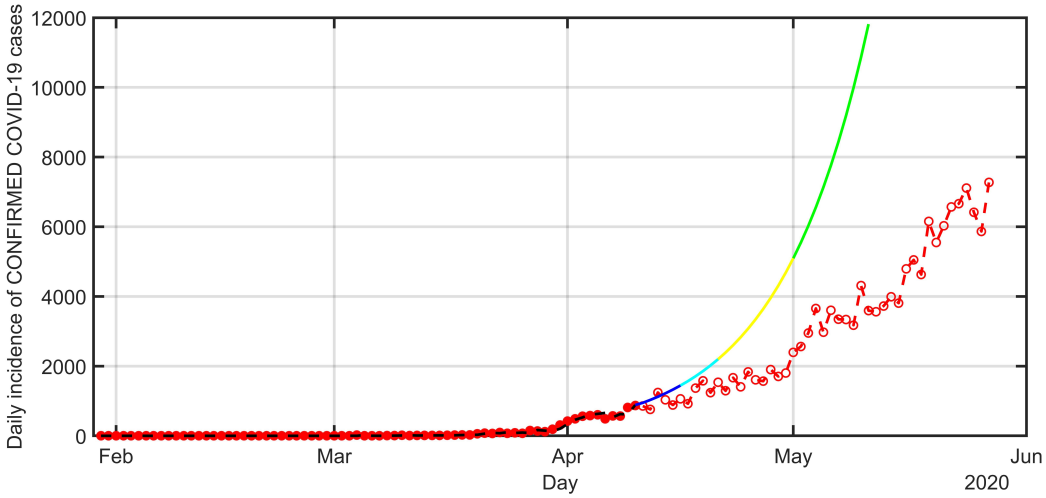

Figure 3

The considered data windows used for parameter estimation overestimated the initial susceptible population, which is more that population of India. The 5 days ahead production is fairly good, whereas the 10, 20 and 30 days ahead prediction is not good. This happens since the data window is very small.

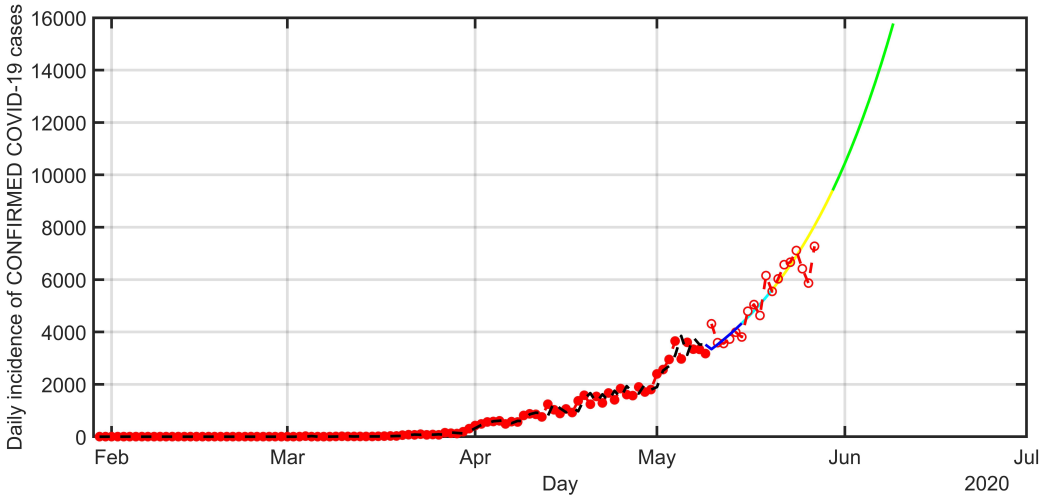

Figure 4

The 5, 10, 15 days ahead prediction of the confirmed cases when data window is 30-Jan-2020 to 9-May-2020 is fairly accurate, since the data window is large this time.

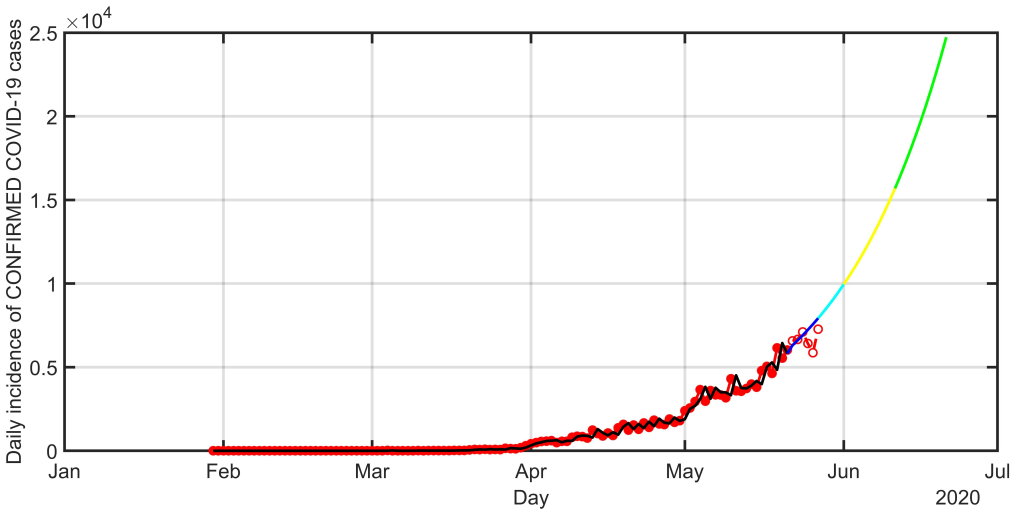

Figure 5

As the data window is expanded, the estimates of both the reproduction number and initial reproduction number stabilises around 1.2 and 1400 million (which is total population of India).

4 Algorithm-2

We start by considering the equation (28), already defined and explained in equation (12)

$$C_{t+1} \sim Binomial\,(S_t, 1-\lambda_{t+1}) \tag{28}$$

Let us define $\lambda_{t+1} = e^{-\beta I_t}$, β is the transmission rate, therefore we have

$$C_{t+1} \sim Binomial\left(S_t, 1-e^{-\beta I_t}\right) \tag{29}$$

In addition to this, we also define :

$$D_{t+1} \sim Binomial\left(I_t, 1-e^{-\gamma}\right) \tag{30}$$

For sake of completeness, we again define the recursion relations which are

$$S_{t+1} = S_t - C_t \tag{31}$$

$$I_{t+1} = I_t + C_{t+1} - D_{t+1}. \tag{32}$$

For large S_t and I_t, the binomial distribution can be approximated by Poisson distributions, i.e. from equation (28) we have

$$\lim_{s_t\to\infty} P(C_{t+1}=c_{t+1}\,|C_t=c_t,S_t=s_t)=\lim_{s_t\to\infty}\begin{pmatrix} s_t \\ c_{t+1}\end{pmatrix}\left(\frac{\lambda}{s_t}\right)^{c_t}\left(1-\frac{\lambda}{s_t}\right)^{s_t-c_{t+1}} \tag{33}$$

$$=\frac{\lambda^{c_{t+1}}e^{-\lambda}}{c_{t+1}!}. \tag{34}$$

Where,

$$(1-\lambda_{t+1})=\frac{\lambda}{s_t} \tag{35}$$

Substituting $\lambda_{t+1}=e^{-\beta I_t}$ in equation (35) we get

$$\lambda=s_t\left(1-e^{-\beta i_t}\right). \tag{36}$$

Approximating $e^{-\beta i_t}=1-\beta i_t$, and substituting in equation (36), we get

$$\lambda=\beta s_t i_t. \tag{37}$$

The term $s_t i_t$ corresponds to homogeneous mixing of susceptible population and infectives.

Therefore, we have

$$P\left(C_{t+1}=c_{t+1}\,|C_t=c_t,S_t=s_t:\lambda=\beta s_t i_t\right)=\frac{\left(\beta s_t i_t\right)^{c_{t+1}}e^{-\beta s_t i_t}}{c_{t+1}!}. \tag{38}$$

Similarly, equation (30) is approximated to Poisson distribution, which is given as

$$P\left(D_{t+1}=d_{t+1}\,|D_t=d_t,S_t=s_t,I_t=i_i:\lambda=\gamma i_t\right)=\frac{\left(\gamma i_t\right)^{d_{t+1}}e^{-\gamma i_t}}{d_{t+1}!}. \tag{39}$$

Here, γ is removal rate.

Therefore, one obtains

$$\begin{aligned} C_t &\sim Po\left(\beta S_t I_t\right) \\ D_t &\sim Po\left(\gamma I_t\right) \end{aligned} \tag{40}$$

If S_t and I_t are known at each point in time $t=1,2,\ldots,T$, estimates for β and γ can be found using maximum likelihood.

4.1 Statistical Estimation of β and γ

The estimates of the unknown parameters β and γ given

- initial susceptible population $S_0 = s_0$,
- initial number of infectives $I_0 = i_0$
- incidences of new generations of infectives $C_0 = c_0, C_1 = c_1, \ldots, C_t = c_t$
- incidences of new death and recovery $D_0 = d_0, D_1 = d_1, \ldots, D_t = d_t$

can be calculated using maximum likelihood estimation procedure.
The likelihood function for estimating β is defined as:

$$L\left(\beta \,|\, S_0 = s_0, I_0 = i_0, C_0 = c_0, C_1 = c_1, \ldots, C_t = c_t\right) = \prod_{k=0}^{T} \frac{\left(\beta s_k c_k\right)^{c_k} e^{(-\beta s_k c_k)}}{c_k!} \tag{41}$$

we use recursive relations defined in equation (31) and equation(32) in equation (41). We maximize the negative log of equation (41), to obtain estimate β.
The estimate of β is

$$\hat{\beta} = \frac{\sum_{k=0}^{T} c_k}{\sum_{k=0}^{T} s_k i_k} \tag{42}$$

Similarly, we define likelihood function for estimating parameter γ, given as

$$L\left(\gamma \,|\, S_0 = s_0, I_0 = i_0, C_0 = c_0, C_1 = c_1, \ldots, C_t = c_t\right) = \prod_{k=0}^{T} \frac{\left(\gamma i_k\right)^{d_k} e^{(-\gamma i_k)}}{d_k!} \tag{43}$$

we use recursive relations defined in equation (31) and equation(32) in equation (43). We maximize the negative log of equation (43), to obtain estimate γ.
The estimate of γ is

$$\hat{\gamma} = \frac{\sum_{k=0}^{T} d_k}{\sum_{i=1}^{T} i_k}. \tag{44}$$

In many epidemiological studies it has been shown that akin to General linear model (GLM) (simple linear regression), distribution model coming from exponential family can

be expressed as Generalised linear model(GLIM). Using the concepts of GLiM, we express Equation (40) as log-link Poisson GLM with offset $\log(s_t i_t)$ and intercept $log(\beta)$.
This is explained in detail in further sections with some basics.

4.2 Basics of Generalised Linear Model

The structural form of a statistical model characterizes the patterns of interactions and associations between various explanatory variable, whereas the model parameters provide measures of strength of associations between these variables.
The **General linear model (GLM)** is a linear statistical model, which is expressed as

$$Y = X\theta + E. \tag{45}$$

Here, $Y \in \Re^{m\times k}$ is a response variable matrix with k response variable. $X \in \Re^{m\times l}$ is explanatory variable matrix with l independent variable. $E \in \Re^{m\times k}$ is noise matrix. The errors $E_i \in \Re^m$ are assumed to be uncorrelated with measurements $Y_i \in \Re^m$, and follow a multivariate normal distribution.
With this, the objective is to model the expected value of Y as a linear function of explanatory variable

$$E[Y_i] = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_l^T \end{bmatrix} \theta_i, \forall i = 1,2,\ldots,k \tag{46}$$

$X_i^T, \forall i = 1,2,\ldots,l$ represent the i^{th} row of matrix X.

The model structure is given as:

$$Y_i = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_l^T \end{bmatrix} \theta_i + E_i, \forall i = 1,2,\ldots,k \tag{47}$$

Here, $Y_i \in \Re^m$ and $\theta_i \in \Re^m$.

$Y_i \in \Re^m$ and E_{ij} are assumed to be normally distributed, such that $E_{ij} \sim N(0,\sigma^2)$

and $Y_i \sim N\left(\begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_l^T \end{bmatrix} \theta_i, \sigma^2 \mathbb{1}\right)$. We estimate the parameter matrix θ using appropriate estimation methods.

The term generalized linear model ($GLiM$) refers to a larger class of statistical models popularized by McCullagh and Nelder. In these models, the response variable Y_i is assumed to follow an **exponential family distribution** with mean μ_i, which is often some nonlinear function of explanatory variable. But McCullagh and Nelder considered μ_i to be a linear function of explanatory variable, therefore we again have:

$$\mu_i = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_l^T \end{bmatrix} \theta_i. \tag{48}$$

But, remember that Y_i comes from a family of exponential distribution like Poisson distribution, normal distribution, binomial distribution.

4.3 Problem Formulation

Consider equation (38), i.e.

$$P\left(C_{t+1} = c_{t+1} | C_t = c_t, S_t = s_t : \lambda = \beta s_t i_t\right) = \frac{\left(\beta s_t i_t\right)^{c_{t+1}} e^{-\beta s_t i_t}}{c_{t+1}!}. \tag{49}$$

Any singe parameter distribution from exponential family can be transformed into the following form:

$$f_X\left(X = x | \theta\right) = \exp\left[a\left(x\right) b\left(\theta\right) + c\left(\theta\right) + d\left(x\right)\right] \tag{50}$$

where $a(x)$, $b(\theta)$, $c(\theta)$ and $d(x)$ are known functions. Similarly, we transform equation (49), we get:

$$P_{C_{t+1}}\left(C_{t+1} = c_{t+1} \middle| \lambda_t = \beta s_t i_t\right) = \exp\left\{\log\left[\frac{\left(\beta s_t i_t\right)^{c_{t+1}} \exp\left(-\beta s_t i_t\right)}{c_{t+1}!}\right]\right\} \tag{51}$$

$$= \exp\left\{c_{t+1} \log\left(\beta s_t i_t\right) - \beta s_t i_t - \log\left(c_{t+1}!\right)\right\}. \tag{52}$$

Let us define a log link function η, given as

$$\eta = \log \lambda \tag{53}$$

the name log link is justified.
From equation (37) and (53) we get,

$$\eta = \log \lambda = \log (E[C_{t+1}]) = \log \left(\beta s_t i_t\right) \tag{54}$$

Further, simplifying equation (54), we get

$$= \log \beta + log\,(s_t i_t)\,. \tag{55}$$

We can assume that log link function $\eta = \log (E[C_{t+1}])$ can depend on more explanatory variable, therefore we can express equation (55) as:

$$\log (E[C_{t+1}]) = \log \beta + \log (s_t i_t) + \alpha_1 X_1 + \cdots + \alpha_l X_l \tag{56}$$

Where, $X_i, \forall i = 1,2,\ldots k$ are feasible explanatory variables.

In this work, we have just considered equation (55) with non-homogeneous mixing of susceptible and infectives i.e. $\lambda = \beta s_t^{\alpha_1} i_t^{\alpha_2}$.
For non homogeneous mixing, we modify equation (55) as

$$\log (E[C_{t+1}]) = \log \beta + \log \left(s_t^{\alpha_1} i_t^{\alpha_2}\right) \tag{57}$$

For, $D(t)$ in equation (40), we can follow the same procedure and link many explanatory variable like age demography of the population, immunity structure of the population etc. Due to the lack of available data for these variables. We simply go with (from equation (40)):

$$E[D_{t+1}] = \gamma i_t \tag{58}$$

4.4 Estimation of GLiM parameters

The estimates of the unknown parameters α_1 and α_2 given

- Initial susceptible population $S_0 = s_0$.
- Initial number of infectives $I_0 = i_0$.

- Incidences of new generations of infectives $C0 = c_0, C1 = c_1, ..., C_t = c_t$

can be calculated using maximum likelihood estimation procedure.
The likelihood function for estimating α_1 and α_2 is defined as:

$$L\left(\alpha_1, \alpha_2 \middle| S_0 = s_0, I_0 = i_{0,} C_0 = c_0, C_1 = c_1, \ldots, C_t = c_t\right) = \prod_{k=0}^{t} \left[c_i \log\left(\beta s_k i_k\right) - \beta s_k i_k - \log c_i!\right]. \tag{59}$$

We use recursive relations defined in equation (31) and equation(32) in equation (59). We maximize the negative log of equation (59), to obtain estimates of α_1 and α_2.
Since, parameters appear in non-affine form we need to use Newton Raphson method for calculating parameters.

4.5 Results

Data Window (data used for prediction of parameters)	α_1	α_2	β	γ	
30-Jan-2020 to 9-April-2020	0.7	1.1	2.3454×10^{-8}	0.0199	Figure-6
30-Jan-2020 to 9-May-2020	0.4	0.9	5.4709×10^{-5}	0.0338	Figure-7
30-Jan-2020 to 21-May-2020	0.6	1	2.9025×10^{-7}	0.0415	Figure-8

In Figure 6, 7 and 8, the **filled** dots represent the data window i.e. the data used for estimating the parameters. The **unfilled** dots represents the observed data. The **black dotted** line is the estimated curve.
In all figures, blue color represents 5 days ahead prediction starting from the end of data window, cyan color represents 10 days ahead prediction starting from the end of data window, yellow color represents 20 days ahead prediction starting from the end of data window, green color represents 30 days ahead prediction starting from the end of data window.
Daily incidence of confirmed COVID-19 cases

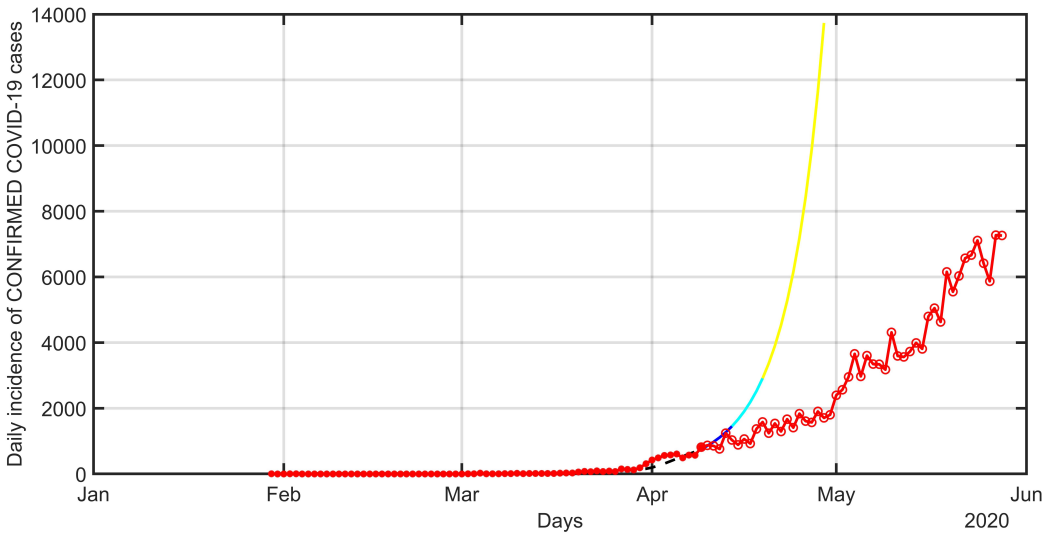

Figure 6

Again, when the data window is small, the resultant parameter estimates are not good. It can be seen that 5 days ahead prediction is good. After, which daily incidence of confirmed cases explodes and moves away from the observed value. When the data

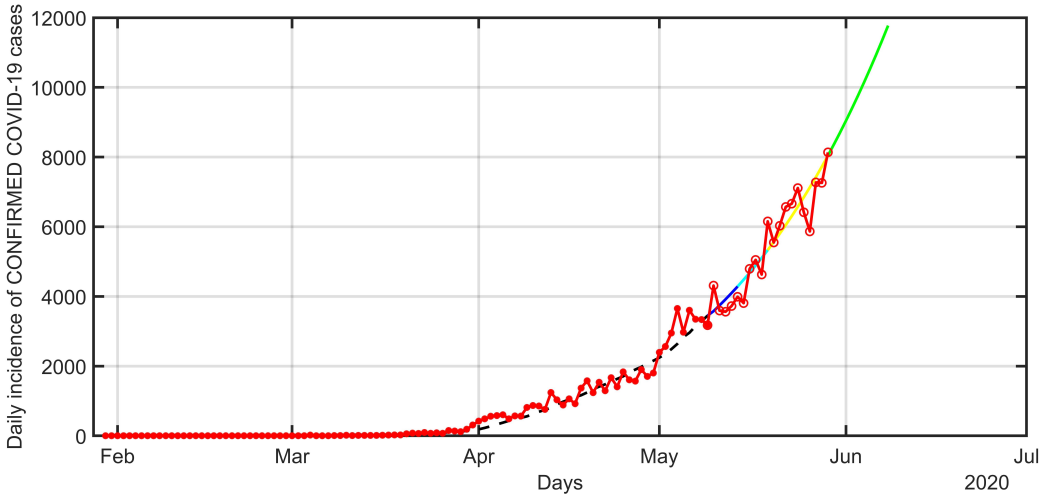

Figure 7

window is large, the parameter estimates are good and leads to fairly accurate 5, 10, 20 days ahead prediction.

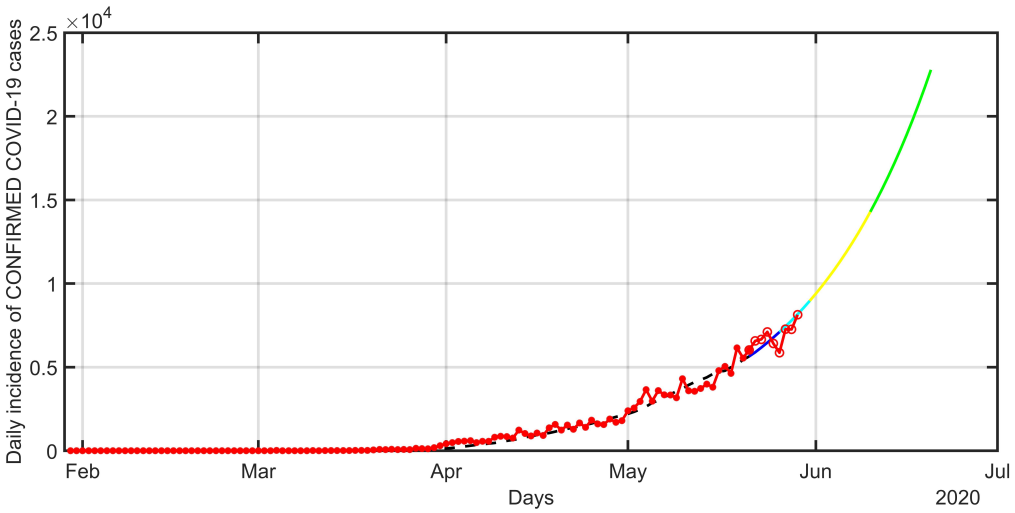

Figure 8

Daily incidence of COVID-19 Recovered + Death cases

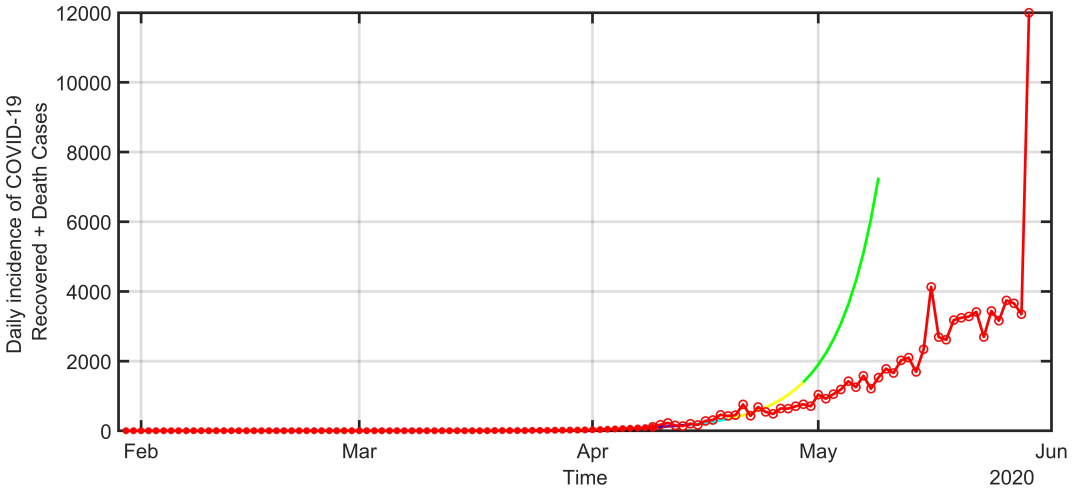

Figure 9: Data window:30-Jan-2020 – 9-April-2020

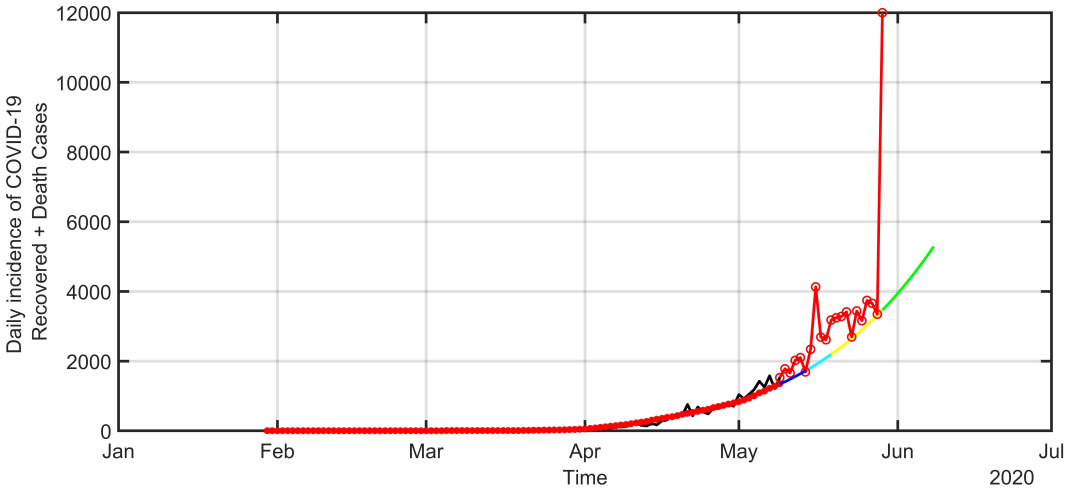

Figure 10: Data window:30-Jan-2020 – 9-May-2020

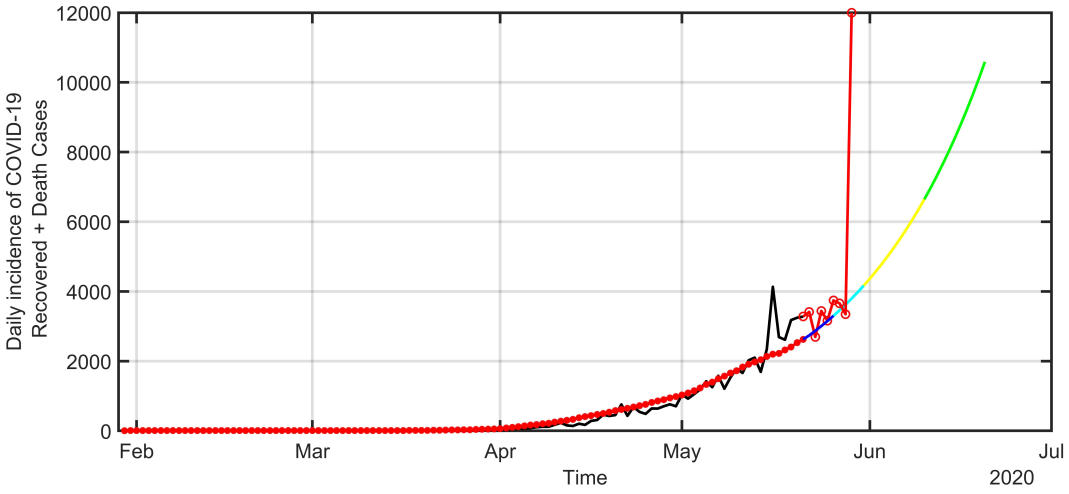

Figure 11: Data window:30-Jan-2020 – 21-May-2020

Cumulative confirmed COVID-19 cases

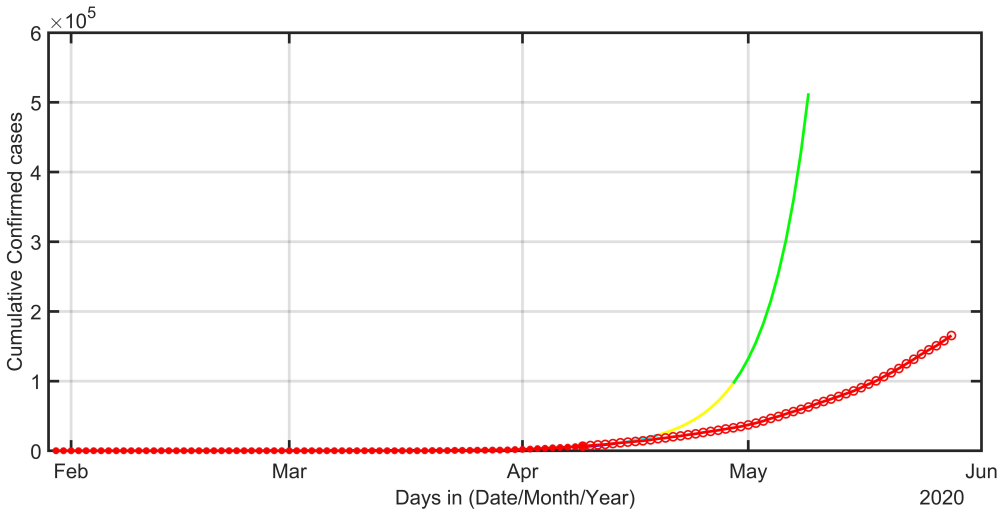

Figure 12: Data window:30-Jan-2020 – 9th-April-2020

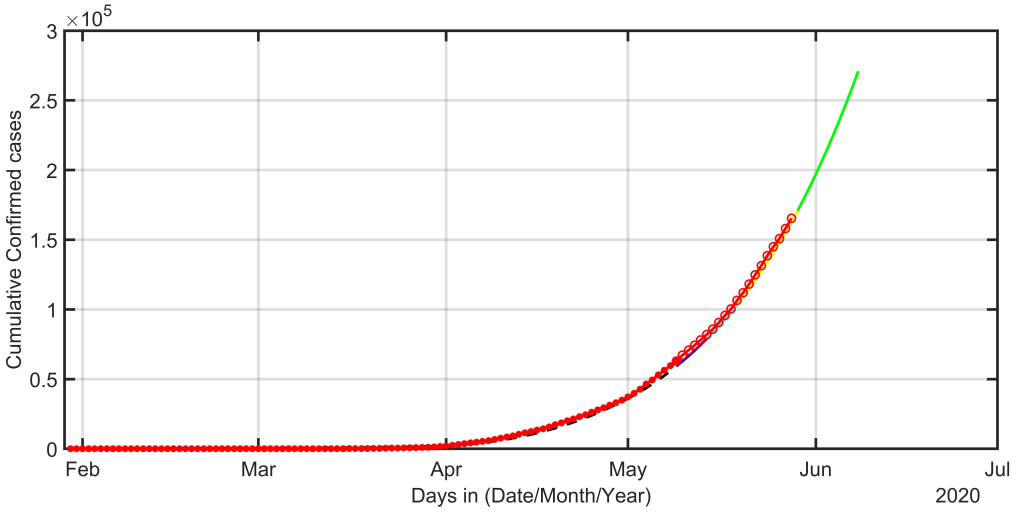

Figure 13: Data window:30-Jan-2020 – 9-May-2020

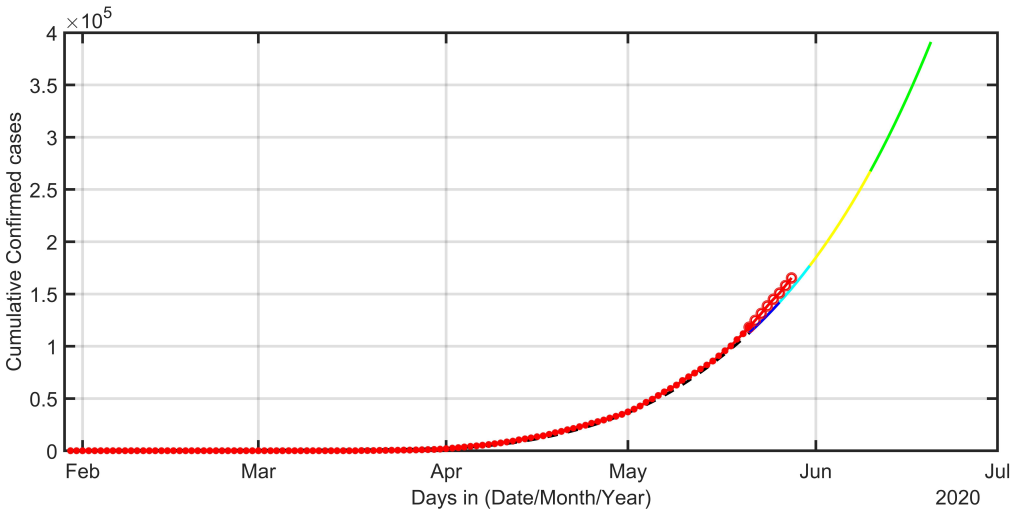

Figure 14: Data window:30-Jan-2020 – 21-May-2020

Cumulative Recovered + death cases

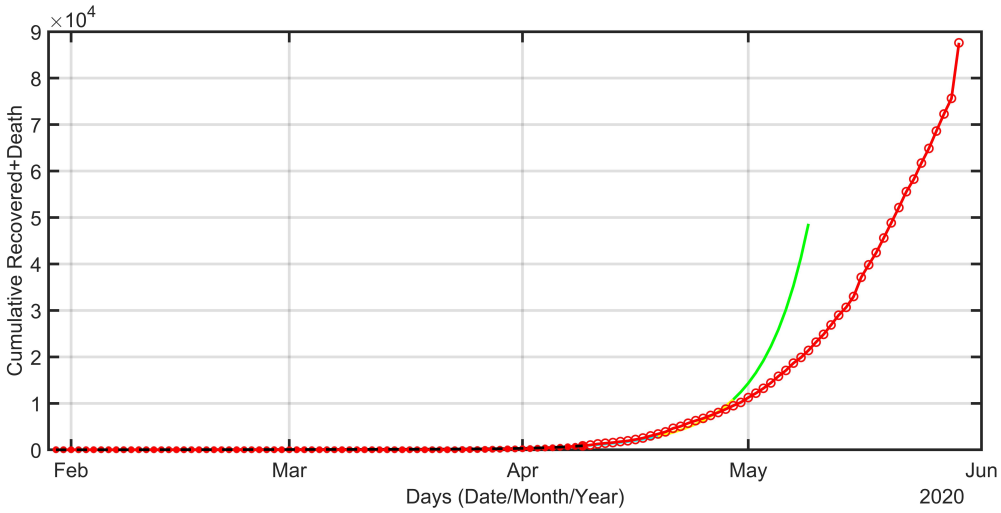

Figure 15: Data window:30-Jan-2020 – 9-April-2020

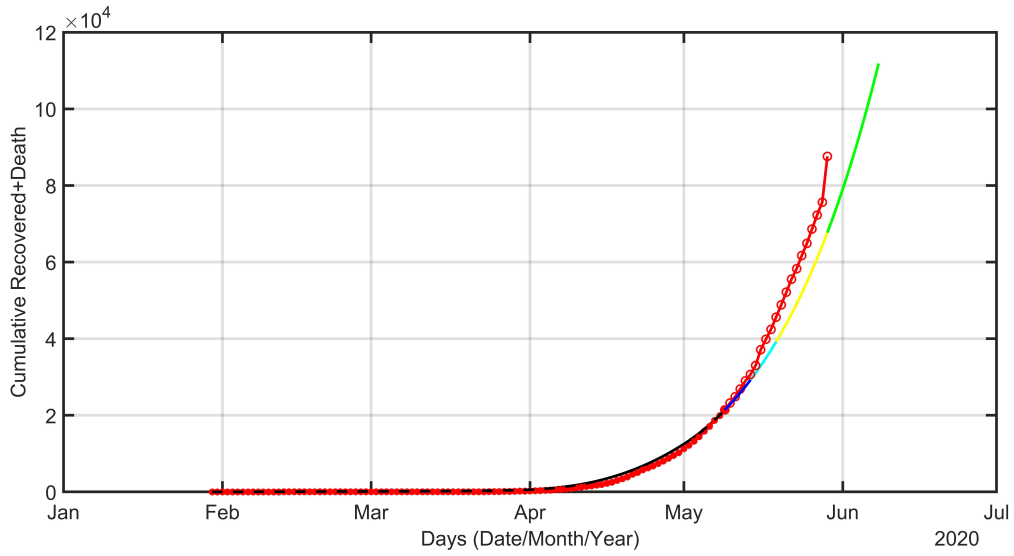

Figure 16: Data window:30-Jan-2020 – 9-May-2020

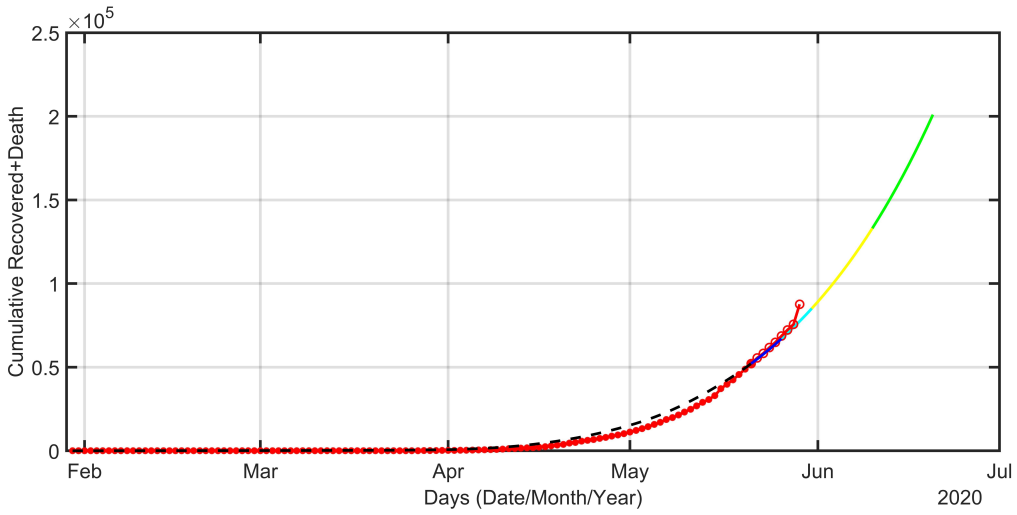

Figure 17: Data window:30-Jan-2020 – 21-May-2020