COVID -19 Analysis and prediction using SIR model

Kunal Gandhi, Prathamesh Thakur, Sanika Patil. (Students, Department of Information Technology, Terna Engineering College Nerul, Navi Mumbai, India)

Prof. Dr. Vaishali Khairnar
(Guide, Department of Information Technology, Terna Engineering College Nerul, Navi Mumbai, India)

Abstract— **As we are aware that Corona Virus is on a rampage across the world and is a major concern at present as it has affected exponential number of people all across the globe and has burgeoned within a very short span of time. Thus, pandemic simulation is essential as a scenario simulation and is performed by several kinds of models, the classical differential models, SIR model and the SEIR model are among the several models that are used to predict and study the outspread of a pandemic; the SIR model being the fundamental model and basic model to be used for the same.**

Keywords: Covid-19, infected, recovered, reproduction rate, recovery rate, susceptible, transmission rate.

I.) INTRODUCTION

COVID -19 is growing fast in many countries around the world. In the absence of availability of a proper medicine or vaccine, currently social distancing, self-quarantine and wearing a face mask have been emerged as the most widely-used strategy for the mitigation and control of the pandemic. In this context, mathematical models are required to estimate disease transmission, recovery, deaths and other significant parameters separately for various countries. In order to get an understanding of the outspread and predicting the further outspread of the virus, the SIR model will help us to visualise and predict the future possible outcome for the same. It also assists us to comprehend the virus' spread in different countries, especially India, taking into consideration number of people that are susceptible, infected and those who have recovered or died from this virus.

II.) LITERATURE SURVEY

A. *Existing System*

COVID-19 cases prediction using linear regression algorithms will assist to predict the future cases of the disease with an approximate number of total people infected and total number of deaths on the basis of the current data which will be processed by the linear regression algorithm, this is not as accurate in prediction with success of up to 56-62 percent. It does not take into consideration the flexibility and extreme variations in the uprise or the downfall of the cases over the period of time as it is evident in our present time because the fit of the curve of the data is not cohesive nor coherent. Other machine learning approaches though more accurate, does not account the flexibility and ease of implementation of the compartmental SIR model for epidemiology.

Major issues associated with the existing system:

1.) Accuracy percentage is on the lower side of the accuracy spectrum since fit to the curve is based on the present data, which may burgeon with passing time leading to contingencies. 2.) To concisely comprehend the behaviour of the COVID-19 virus with machine learning approaches, the implementation is time consuming and inculcates several unnecessary parameters which eventually turn into obscurity.

B. *PROPOSED SYSTEM*

Just because the rise in number of cases is exponential, it does not imply that we can fit the data to an exponential curve and predict the number of cases in the coming days. Compartmental model techniques are normally used to model infectious diseases. Same could be used in the case of COVID-19 too.

The main aim of our system is to predict the behaviour of the COVID-19 virus as to how it is going to affect the human race in upcoming days. SIR model is a compartmental model used for simplifying mathematical modelling of infectious disease. The compartments that will be made of the total population will be:

- S: Susceptible (can potentially catch the disease)
- I: Infected (have disease and can spread it.)
- R: Removed (who caught the disease but have recovered or died.)

We will have mathematically conducted the analysis and would be able to predict:

- The total number of people who might get infected by the COVID-19 virus.
- How many people may catch the disease?
- How many deaths we may witness?

Using the compartmental SIR model for epidemiology, a concise and streamlined approach is implemented which not only predicts with accuracy but also considers the erratic fluctuations in the curve which occur due to irregular uprise or downfall of the number of cases, with the assistance of designated parameters, Fig. 6:

- *Beta* (*ß*): Average number of contacts per person per time (transmission between S and I)
- *Gamma (Γ)*: Rate of recovery or mortality rate = Number of recovered/ Number of infected
- *Ro*: An estimate of the number of people an average infected person will spread the disease to.

This will enable us to know the future statistics of the disease and assist us to take steps in order to prevent or at least take precautions to avoid the spreading of the pandemic in the world and bolster our efforts to not reaching the expected or the predicted numbers as it may bring worst case scenarios in the world. We can then make predictions for its behaviour for individual countries, in our case, India.

III.) METHODOLOGY

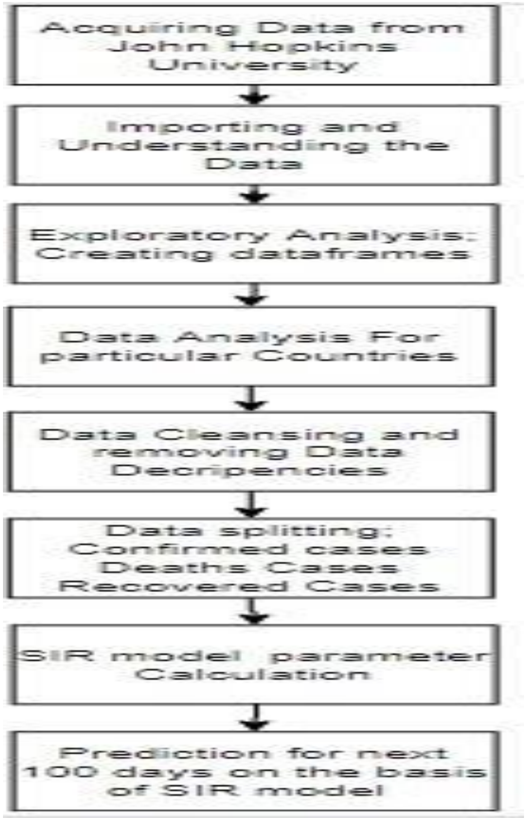
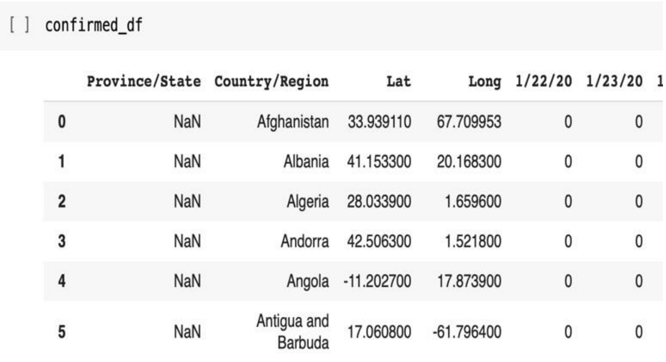


Fig.1

A. *Explanation:*

This will be implemented using python code, concisely by using “scipy” library.

1. Getting raw dataset from John Hopkins university. The data is split across the following three files
- Deaths
- Recovered
- Global cases.
2. Using pandas, the dataset can be directly imported into data-frames, Fig. 2. It is much better to use the URLs than to download the file manually and then read it as it becomes easier to load and refresh the analysis with new data.

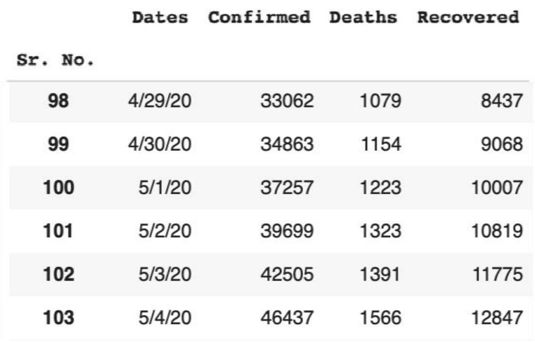


```
[ ] confirmed_df
```

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1
0	NaN	Afghanistan	33.939110	67.709953	0	0	
1	NaN	Albania	41.153300	20.168300	0	0	
2	NaN	Algeria	28.033900	1.659600	0	0	
3	NaN	Andorra	42.506300	1.521800	0	0	
4	NaN	Angola	-11.202700	17.873900	0	0	
5	NaN	Antigua and Barbuda	17.060800	-61.796400	0	0	

Fig.2

3. First, some exploratory analysis on the data is conducted and summarising of some stats and comprehending the trend in countries like India, United States of America and Spain is done. (Fig. 3, Fig. 4, Fig. 5 respectively).



Sr. No.	Dates	Confirmed	Deaths	Recovered
98	4/29/20	33062	1079	8437
99	4/30/20	34863	1154	9068
100	5/1/20	37257	1223	10007
101	5/2/20	39699	1323	10819
102	5/3/20	42505	1391	11775
103	5/4/20	46437	1566	12847

Fig. 3: Cases in India

Sr. No.	Dates	Confirmed	Deaths	Recovered
116	5/17/20	1495736	92983	272265
117	5/18/20	1518126	94210	283178
118	5/19/20	1539133	95666	289392
119	5/20/20	1561830	97167	294312
120	5/21/20	1587596	98352	298418
121	5/22/20	1611253	99569	350135
122	5/23/20	1632364	100658	361239

Fig. 4: Cases in USA

Sr. No.	Dates	Confirmed	Deaths	Recovered
50	3/12/20	2277	55	183
51	3/13/20	5232	133	193
52	3/14/20	6391	195	517
53	3/15/20	7798	289	517
54	3/16/20	9942	342	530
55	3/17/20	11748	533	1028
56	3/18/20	13910	623	1081

Fig. 5: Cases in Spain

4. Removing data discrepancies where data is not updated or has null values in it. Once these discrepancies are removed, forming data frames comprising of total number of cases, deaths and recoveries, for ease of understanding the trend of the virus.
5. Then we will model the data on the SIR epidemic model using its designated parameters; beta, gamma and Ro, using mathematical approach.
6. The model consists of three compartments: S for the number of susceptible, I for the number of infections, and R for the number of recovered or deceased (or immune) individuals. This model is reasonably predictive for infectious diseases which are transmitted from human to human, and where recovery confers lasting resistance.

Each member of the population typically progresses from susceptible to infectious to recovered. This can be shown as a flow diagram, Fig. 3, in which the boxes represent the different compartments and the arrows the transition between compartments i.e.: Susceptible, Infectious and Recovered.

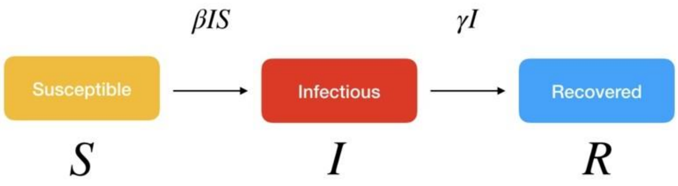


Fig. 6

7. Once the model is created predictions can be made for individual countries.

IV.) IMPLEMENTATION

The SIR model has the following major assumptions,

- Population remains constant.
- Rate of infective(ß) is proportional to contacts.
- Infected either recover/ die at constant rate, Γ.

In the model, the value Ro is an estimate of the number of people an average infected person will spread the disease to. If the value of Ro is greater than 1, then the disease probably continues to spread and if it is < 1 then the disease slowly dies down. Since COVID-19's Ro is > 2, so an average infected person spreads it to 2 or more people who again spread it to 2 or more people and that is how this infection continues to spread across the globe. There are other parameters in the model like and which needs to be estimated. We will implement our SIR model specifically for India. We have three major differential equations for SIR (S -Susceptible, I-Infectious, R- Recovered/dead) model,

$$\frac{S}{t} = -\,ß\,\underline{S}.\ \ldots(q.1).$$

$$\frac{\;}{t} = +ß\ -\ (S.\ ) - \Gamma.\ \mathrm{I}\ldots(\mathrm{eq.}\ 2)$$

$$\frac{R}{t} = \Gamma.\ \ldots(q.3)$$

Where,

ß = Average number of contacts per person per time (transmission between S and I),

Γ = Rate of recovery or mortality rate = Number of recovered/ Number of infected

N = S + I + R initially, which is independent of time 't' and denotes the entire population.

The eq. 1 indicates a direct contact between the susceptible and the infected based on our second assumption at a constant rate ß which is the rate of contact or the transmission rate. Since at the beginning, the entire population is susceptible to the virus, the number of susceptible cases gradually go on decreasing, signified by the negative sign, as they become infected cases.

Consequently in eq. 2, the rate of change of the infected with respect to time, goes on increasing at a constant rate ß, at the same time, based on our third assumption, certain number of infectives either recover or die at a constant rate Γ, which is signified by (-Γ.I) part of the eq.2 having the negative sign.

The gain in eq. 3 suggests that the number of people removed from the Infected category move into the recovered category at a constant rate Γ. We now have our primary differential equations for our SIR model.

B. *Calculating the parameters:*

As pointed out earlier, when there is no infection, I+R=0, then by substituting S≈N into, we obtain the following equation:

$$\frac{\sim}{t}\quad ß\,(-\Gamma\ \ldots)\qquad (q.4\ )$$

Then, by integrating eq. 4 we get the following equation:

$$= o.\ ^{ß-\Gamma\ t}\ldots.\qquad (q.5)$$

Where Io is the number of infectives at the start of the disease spread. We can now determine the values of ß and m. At the onset of the infection, almost the entire population is susceptible: i.e., S≈N. Therefore, I(t) first grows exponentially.

$$\frac{\simeq}{t}\ .\ \ldots\qquad \textit{(eq. 6)}$$

where m=β-γ is a constant term that represents the difference between transmission and recovery rates.

$$(t\ )\ \sim\ o.\ ^{t}\ldots\ (\mathrm{eq.}\ 7)$$

Taking log on both sides we get,

$$n\ = t + n\ o\ldots(q.8)$$

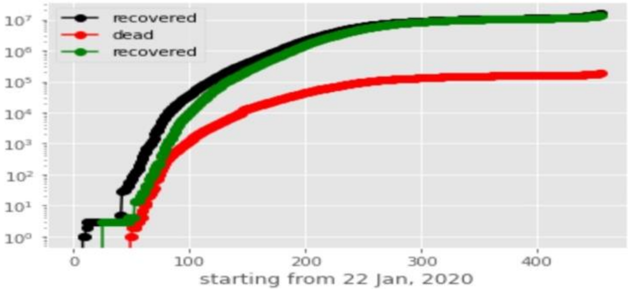


Fig. 7

The m value can be estimated from the log-plot data(fig.9) and by using, for example, least squares to obtain the best- line fit(fig.8).

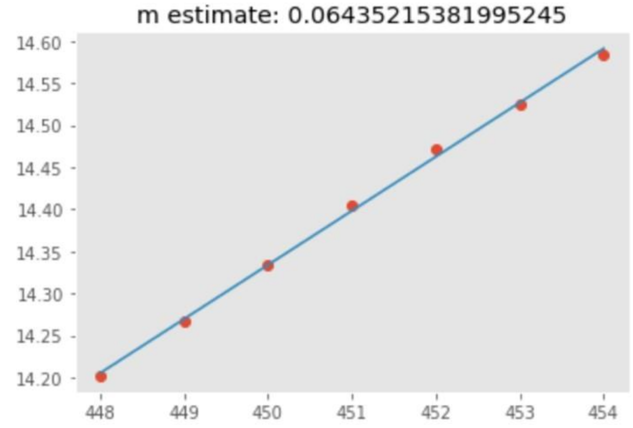


Fig. 8: m-estimate for India

Determining the value of Gamma(Γ):

Suppose I(t)=Io (constant), then we obtain the following equation from eq.3:

$$\frac{R}{t} = \Gamma . o \ldots (q.9)$$

Thus, by integration of eq. 9, we get:

$$R(t) = \Gamma . t . o \ldots (q\ 10)$$

If it takes t=T days to recover, then, R(T)=Io, or Γ.T=1.

Thus, we obtained the following equation:

$$\overset{1}{\Gamma} \approx \ldots q.11 \quad (\qquad)$$

Where T is the time required in days, for the recovery. From eq.3, for a change in time of dt=a, we obtain the following equation:

$$\frac{R(t+)}{} = \Gamma . \ldots \quad (q.12)$$

$$\Gamma \approx \frac{R(t+1) - R(t)}{(\ )} \ldots . \ (q.13)$$

Note that from eq.13 and eq.8 we can obtain the value of Beta or the transmission rate (ß).
By directly estimating from the data set for India we obtain the following values:

```
gamma = '0.08156722521179009' beta = '0.14591937903174254'
Ro = '1.7889462177101336'
```

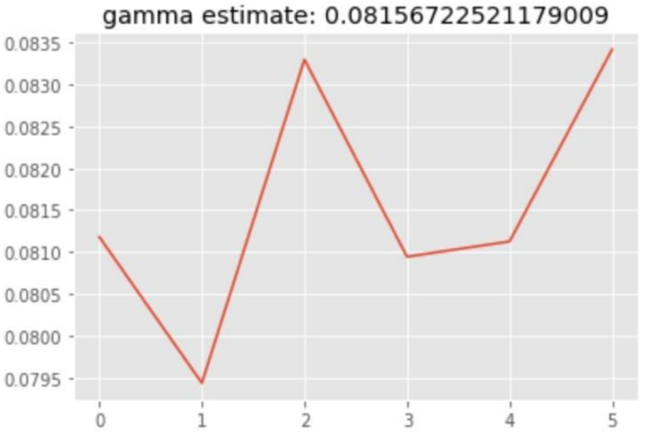
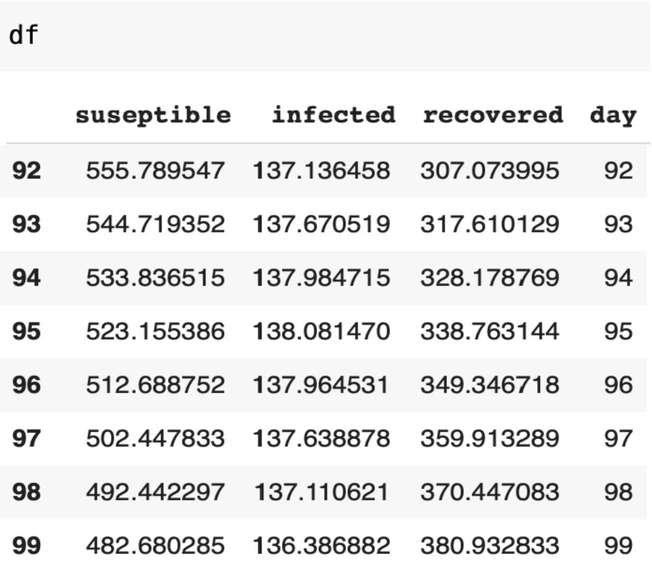


Fig. 9: Gamma or the recovery rate(Γ) for India

In fig 8. gives us our Recovery rate(Γ), Transmission rate(ß) and the Basic Reproductive ratio (Ro) which is an estimate of the number of people on an average, the infected individual will spread the disease to and is calculated by:

$$\overset{ß}{Ro} = - \ldots (q.14) \qquad \Gamma$$

By implementing all the above equations in Python and developing a function for the SIR model which calibrates the number of susceptible, infected and recovered/ dead cases the following result is obtained:



```
df
```

	suseptible	infected	recovered	day
92	555.789547	137.136458	307.073995	92
93	544.719352	137.670519	317.610129	93
94	533.836515	137.984715	328.178769	94
95	523.155386	138.081470	338.763144	95
96	512.688752	137.964531	349.346718	96
97	502.447833	137.638878	359.913289	97
98	492.442297	137.110621	370.447083	98
99	482.680285	136.386882	380.932833	99

Fig 10: Possible number of cases of COVID-19 and their category in the coming 100 days in India.

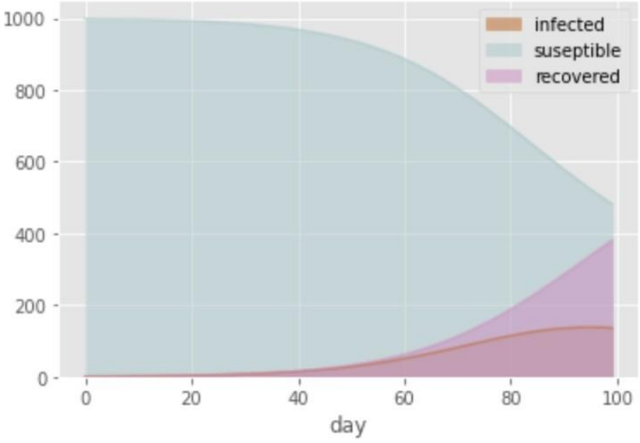


Fig. 11: Graph of the infected, susceptible and recovered cases for the time span of 100 days in India

The Fig.10 and Fig.11 suggests, in the coming days the number of susceptible cases will gradually decrease but at the same time due the basic reproductive ratio (R≈ 1.8), from fig. 9 being greater than 1, the disease will spread. Thus, the number of Infectives will exponentially increase at the same time. The recovery rate on the other hand compared to the exponential rise of infected individuals is on the lower side due the average recovery period being fourteen days for any infected individual.

V.) CONCLUSION

In our study we have taken the entire Covid- 19 dataset from John Hopkins University Centre for System Science and Engineering (JHU-CSSE), have simulated the SIR model on the Google Collab platform and have estimated the possible number of cases in the coming days. Based on our initial assumptions and the results thus obtained though not exceptionally accurate, show otherwise that the necessity for a lockdown and restricted movement is imperative to reduce the number of direct contacts as there will be a sharp increase in the number of infectives and comparatively due to a slower rate of recovery the disease will remain active for a prolonged period of time.

VI.) REFERENCES

1.) https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model
2.) https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology
3.) https://www.youtube.com/watch?v=NKMHhm2Zbkw&t=336s
4.) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7388782/
5.) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7570398/
6.) https://ddd.uab.cat/record/108432
7.) https://link.springer.com/article/10.1186/s13662-017-1078-5