Analyzing Data on the Spread of COVID-19 Using Statistical Tools to Predict the Inflexion Point of the Virus in Italy

** Nirbhay Narang*
*** Mehul Jangir*

Abstract

The outbreak of the novel coronavirus COVID-19 had resulted in deaths of over 24,000 people by April 20, 2020. The goal of this paper is to use and apply principles of statistics and machine learning on COVID-19 datasets available online to predict the inflection point of the spread of the virus. The inflection point, for the purpose of this paper, is defined as a point in time in days after the outbreak of the virus at which there is a change in the direction of the rate of spreading of the virus. We are using available libraries to fit the data a logistic function.

Keywords : **Coronavirus, curve-fitting, Logistic functions, inflection points, infection prediction, Python**

I. LOGISTIC FUNCTIONS

Logistic functions have been commonly used to describe the growths of populations [1]. Considering a viral infection to be the growth of the population of a pathogen, it is rational to fit a logistic function to our data.

The general equation for a logistic function is

$$f(x,a,b,c) = \frac{c}{1+e^{\frac{-(x-b)}{a}}} \qquad (1)$$

In (1), the variable x is defined as the time in days after the outbreak. The constant a refers to the infection speed (rate of infection) of the virus. The constant c denotes the total recorded number of infected people at the infection's end (in our case, the last day of recorded data), and lastly, the constant b is defined as the day with the highest number of recorded cases. That is, it is the point at which the first derivative starts to decrease. In this context, it can be defined as the peak point after which the outbreak starts to become less aggressive and decreases. As time progresses indefinitely, we can see that the function approaches the value of c - at which point we can conclude that the infection has ended.

$$\lim_{x\to\infty} f(x) = c \qquad (2)$$

II. DATASET

The Italian Civil Protection Department refreshes the cumulative data of infected people daily. This data is publicly available as open data on GitHub. The link to the dataset can be found in the references section of this paper [3].

III. FINDINGS

Using the *scipy.optimize* library on the dataset above to obtain the best-fit values for *a, b, and c* [2], we concluded that the inflection point for Italy based on the data in the dataset, will occur approximately 85 days after *January 1st, 2020.*

Fig. 1 reflects the values for *a,b, and c* being

a = 6.62760507938102 (speed of infection)

Manuscript Received: April 25, 2020; Revised: May 10, 2020; Accepted: May 14, 2020.
*N. Narang is a student of Jayshree Periwal International School, Mahapura Rd, Narayan-Y-Block, Mahapura, Rajasthan, India - 302 026. (email: snirbhay799@gmail.com)
**M. Jangir is a student of Jayshree Periwal International School, Mahapura Rd, Narayan-Y-Block, Mahapura, Rajasthan, India - 302 026. (email: mehuljangir@gmail.com)
Doi :10.17010/ijcs/2020/v5/i2&3/152206

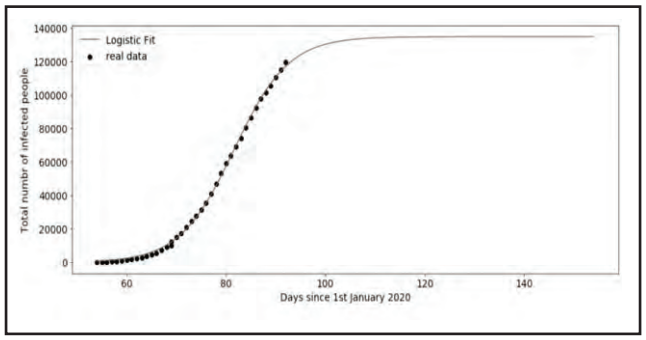

Fig. 1. Total Number of Infected People Against Days Since the Start of the Outbreak

Note. Both the actual number of cases and output of logistic regressor have been plotted.

b = 84.51551700709867 (inflection point in days)

c = 163442.9559024376 (total number of infected people at the end of the infection)

IV. CODE

```
import pandas as pd
import numpy as np
from datetime import datetime, timedelta
from sklearn.metrics import mean_squared_error
from scipy.optimize import curve_fit
from scipy.optimize import fsolve
import matplotlib.pyplot as plt
%matplotlib inline

url = «https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv"
df = pd.read_csv(url)
df = df.loc[:,['data','totale_casi']]
date = pd.to_datetime(df['data']) - pd.to_datetime("2020-01-01T18:00:00")
dt2 = list()
for d in date:
dt2.append(d.days)
df['data'] = dt2
df.head()

def logistic_model(x, a, b, c):
return c / (1+np.exp(-(x-b)/a))

x = list(df.iloc[:,0])
y = list(df.iloc[:, 1])
fit = curve_fit(logistic_model, x, y, p0=[2, 100, 20000])
a = fit[0][0]
b = fit[0][1]
c = fit[0][2]
print(a,b,c)

errors = [np.sqrt(fit[1][i][i]) for i in [0, 1, 2]]
sol = int(fsolve(lambda x: logistic_model(x, a, b, c) - int(c), b))
pred_x = list(range(max(x), sol))
plt.rcParams['figure.figsize'] = [15, 6]
plt.rc('font', size=14)
plt.scatter(
  x,
  y,
  label='real data',
  color='black'
)

#logistic curve
plt.plot(
  x+pred_x,
   [logistic_model(i, fit[0][0], fit[0][1], fit[0][2]) for i in x + pred_x],
  label="Logistic Fit"
)

plt.legend()
plt.xlabel("Days since 1st January 2020")
plt.ylabel("Total numbr of infected people")
plt.show()
```

V. CONCLUSION

In the process outlined above, we used widely available statistical tools and open-datasets to apply machine-learning principles to the spread of the novel coronavirus. We conclude that the inflection point of the virus in Italy should occur sometime around 85 days after January 1, 2020 which is Thursday, March 26, 2020. Our findings also conclude that the total number of recorded cases in Italy will be around *1,60,000.* Comparing this with the number of recorded cases on April 10, 2020 (1,59,516), we can see that our model is somewhat accurate.

REFERENCES

[1] C-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to Logistic Regression analysis and reporting," *J. of Educational Res.*, vol. *96*, no. *1*, pp. 3-14, 2002. Doi: 10.1080/00220670209598786

[2] D. Chakrabarty, "Curve fitting: Step-wise least squares method," *AryaBhatta J. of Mathematics & Informatics,* vol. *6*, no. *1*, pp. 15-25, 2014.
[3] https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv

About the Authors

Nirbhay Narang is a student of Jayshree Periwal International School. He has a keen interest in using technology to help bring about social change. He is striving to develop professional skills in various areas and to build upon his passions for mathematics, developing software, and engineering.

Mehul Jangir is a student at Jayshree Periwal International School. His passions include mathematics, computer programming, astronomy, and engineering. He is a published author. He is eager to be involved with the COVID-19 pandemic and is looking for solutions. He organized an online Model United Nations (MUN) to enable people to discuss COVID-19 and used Python to evaluate the inflexion point for Covid-19 clusters.

INDIAN JOURNAL OF COMPUTER SCIENCE

Statement about ownership and other particulars about the newspaper "Indian Journal of Computer Science" to be published in the 1st issue every year after the last day of February.

FORM 1V
(see Rule 18)

1.	Place of Publication	:	NEW DELHI
2.	Periodicity of Publication	:	BI-MONTHLY
3.	4,5 Printer, Publisher and Editor's Name	:	S. GILANI
4.	Nationality	:	INDIAN
5.	Address	:	Y-21,HAUZ KHAS, NEW DELHI - 16
6.	Newspaper and Address of individual	:	ASSOCIATED MANAGEMENT
	Who owns the newspaper and partner of	:	CONSULTANTS PRIVATE LIMITED
	Shareholder holding more than one percent.	:	Y-21, HAUZ KHAS, NEW DELHI-16

I, S.Gilani, hereby declare that the particulars given above are true to the best of my knowledge and belief.

DATED : March 1, 2020

Sd/-
S. Gilani
Signature of Publisher