Applying Statistical inference techniques to smoothen COVID19 data and predict future infection

ABSTRACT

General Statement

The main purpose of our work is to collect the second phase COVID19 dataset and doing some analysis by mathematical operations or model on it. The analysis is done on Infection and Deaths (Daily and Cumulative). COVID19 data is collected from 15th May 2021 to 15th May 2022 which is one year in duration. In the dataset, there are some errors as it is not continuously updating sometimes or the data may not be correct so we have generated box-plot and find out some abnormal values which is outlier. Then the abnormal values are fixed by using Savitzky–Golay filter. The Savitzky–Golay filter is used to smooth the curve. The smoothen curve is plotted by using the smoothen values. Standard SIRD Model is applied to analyze the number of susceptible, infected, recovered, and deceased people from the COVID19 dataset.

Keywords - second phase COVID19, mathematical operations, box-plot, abnormal values, smooth, SIRD Model

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

As the ongoing pandemic caused by the outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) or simply COVID-19 sweeps the globe at an unprecedented rate, leaving trails of high infection and mortality, it is critical to keep track of the situation. In order to take immediate action, we need to understand the virus's propagation patterns in a host population. To save lives, as well as effective restorative and mitigation measures It has previously been observed in a number of nations. Educating communities that precise and prompt infection testing, tracing, and monitoring result in containment and halting the spread of the disease. The dynamics of growth are examined in this work. Infection in Bangladesh, one of the most vulnerable countries in terms of population density. The pursuit of the highest density in the world is based on reported statistics from March 8, 2020 – January 20, 2021. In this project, the analysis is done on Infection and Deaths (Daily and Cumulative). COVID19 data is collected from 15th May 2021 to 15th May 2022 which is one year in duration. Mainly the reason to do this project is to find out the errors in the data. We won't know the original death number. If we collect long term data and can observe data trend that the data is bigger or smaller and if there is any abnormal data we can identify and correct it. Thus, it is important. So that we could use the correct picture of Covid-19 and will use those pictures for different inferences and prediction. It is possible to identify outliers and conduct unique corrective steps. For projecting future trends, an innovative curve-fitting approach can be applied with prior data values. All of the analytics visualizations and results can be shared with other users right. We have developed a methodology to find out the outlier inference in the data. Then the abnormal values are fixed by using Savitzky–Golay filter. The smoothen curve is plotted by using the smoothen values. Standard SIRD Model is applied to analyze the number of susceptible, infected, recovered, and deceased people from the COVID19 dataset.

1.1 Motivation

As we know it is difficult to collect Covid-19 data for different variation such as for human error like, a person can make mistake while collecting the data by doing over counting taking more values or down counting by taking less values. Sometime it happens that if the person is died for other reason, it might be counted as Covid death and that number is added.

1.2 Objectives

Infected and Death rate has been collected from 15th May 2021 to 15th May 2022. We have stored the data in an excel file of extension .xlsx. We have visualized the graph on the collected data. The outliers of daily infection and death are identified by generating boxplot. Inference technique (Savgol) filter has been used to smooth graph. Smooth data will be predicted using standard SIRD Model.

1.3 Limitation

We have some errors in daily infection and deaths of our dataset and tried to find out the abnormal values by generating boxplot which are outliers. Outliers can be deleted or correct. The abnormal values increase the error of statistical data which can cause bias or influence. They can also impact the basic statement of regression as well as other statistical models. We tried to fix the outliers by smoothing the curve using savgol filter.

1.4 Data Collection

The data has been collected from several websites. The duration is one year from 15th May 2021 to 15th May 2022. The data has been stored in an excel file of extension .xlsx.

1.5 Model

We have used standard SIRD Model in our project. Standard SIRD Model is the traditional model and it works on daily infection and deaths. There are some other SIRD Model but we did not used it as they are rational models.

1.6 Arrangement

In the first chapter of this paper, I have discussed about the Introduction which includes Motivation, Objectives, Limitation, Data Collection and Model of our project. In the second chapter, I have discussed about the methodology of our project. In the third chapter, I have discussed about the Literature Review which includes the summary of related papers. In the fourth chapter, I have discussed about the Results and Discussion which includes all of the figures that we generate and plot in backend of our project. In the fifth chapter, I have discussed about the conclusion of our project. In the sixth chapter, I have added all of the references of research paper of our related work.

CHAPTER 2

METHODOLOGY

2.1 Introduction

In this chapter the methodology for “Applying Statistical inference techniques to smoothen COVID19 data and predict future infection” has been discussed. In section 2.2, 2.3, 2.4, 2.5 and 2.6 discusses about the box-plot, outliers, Savitzky-Golay (Savgol) Filter, smoothing curve and SIRD Model.

2.2 Boxplot

A boxplot is a standardized method of depicting data distributions using a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can provide information about your outliers and their values. We draw a box from the first to the third quartile in a box plot. At the median, a vertical line runs through the box. Box plots divide the data into parts, each of which contains around 25% of the total data in the set. Box plots are valuable because they provide a visual summary of the data that allows researchers to easily discover mean values, data set dispersion, and skewness. Although boxplots appear basic when compared to a histogram or density plot, they have the advantage of taking up less space, which is beneficial for comparing distributions across multiple groups or datasets. A data point that lies outside the box plot's whiskers is referred to as an outlier. Outside 1.5 times the interquartile range, for example, the top quartile is higher than the lower quartile (Q1 - 1.5 * IQR or Q3 + 1.5 * IQR). For generating boxplot of Daily Infected and Deaths, the following code is used:

```
second120days = dataset2['Infected (Last 24 Hours)']
fig2 = go.Figure()
fig2.add_trace(go.Box(x=second120days, name='', marker_color = 'indianred'))
fig2.update_layout(
   xaxis=dict(title='Values from 12Sept21 - 9Jan22')
)
```

2.3 Outliers

An outlier is a value in a data collection that is significantly greater or smaller than the rest of the values. Outliers might be caused by data entry or experiment measurement errors, sample issues, or natural fluctuation. While exploring or entering data, an error can occur. A typo can accidentally type the wrong value during data entry. An outlier could be a sign of poor data. For instance, data could have been wrongly coded or an experiment could have been done erroneously. We can assess whether a given statistic is an outlier by multiplying the interquartile range (IQR) by 1.5. If we subtract 1.5 x IQR from the first quartile, any data values that are less than this number are considered outliers. From our dataset values, the outliers are recognized from the boxplot.

2.4 Savitzky-Golay (Savgol) Filter

The Savitzky-Golay filter is a low-pass filter for data smoothing. To use it, set the window size, which is the number of points used to calculate the fit, and the order of the polynomial function used to fit the signal as input parameters of the function (as a one-dimensional array). Convolution is accomplished by fitting successive sub-sets of neighboring data points with a low-degree polynomial using the linear least squares approach. When the data points are evenly spaced, an analytical solution to the least-squares equations can be found in the form of a single set of "convolution coefficients" that can be applied to all data sub-sets to produce smoothed signal estimates (or derivatives of the smoothed signal) at the central point of each sub-set. Abraham Savitzky and Marcel J. E. Golay popularized the method, which is based on well-established mathematical procedures, by publishing tables of convolution coefficients for various polynomials and sub-set sizes in 1964. The tables have had some mistakes rectified. The approach has been improved to handle two and three-dimensional data. A typical method for smoothing data is Savitzky–Golay (SG) filtering, which is based on local least-squares fitting of the data by polynomials. For applying Savitzky–Golay filter on graph, the following code is used:

```
x1 = dataset['Reporting Date']
y1 = dataset['Infected (Last 24 Hours)']
yhat1 = savgol_filter(y1, 51, 4) # window size 51, polynomial order 4
plt.plot(x1,yhat1, color='red') # plotafter applying savgol filter on it
plt.show()
```

2.5 Smoothing Curve

In order to smooth a data set, we need to use a filter which is a mathematical process that allows us to remove the oscillations caused by the intrinsic noise in our data collection. Python has a number of filters that change based on the mathematical approach used to process the data. A smooth curve is a smooth function, with the word "curve" understood in the context of analytic geometry. A smooth curve, in particular, is a continuous map from a one-dimensional space to a dimensional space with continuous derivatives up to a chosen order on its domain. Smoothing is a technique for decreasing noise in datasets.

2.6 SIRD Model

Mathematical modeling, according to Haines and Crounch, is a process in which real-life situations and relationships are expressed using mathematics. In simplest terms, mathematical modeling is the practice of using mathematics to describe systems (activities). The technique of applying mathematics to model real-world processes and events is known as mathematical modeling. SIRD stands for Susceptible, Infected, Recovered, and Deceased. The SIRD model is a variation on the SIR model that includes two additional assumptions: recovery with immunity and death. The most basic SIRD model involves three sorts of individuals:

(S) Susceptible—the people who could become infected.
(I) Infected—the people who are infected at that moment.
(R) Recovered—the people who have had the disease and are now healthy.
(D) Deceased—the people who have died of the disease.

If we assume that the population affected by the epidemic is N, satisfying N = S + I + R + D, the model, which we will call SIRD, it is governed by the following system of differential equations:

$$\mathbf{S'(t) = -\beta\, S(t)I(t)} \qquad \mathbf{(1)}$$

$$\mathbf{I'(t) = \beta\, S(t)I(t) - \alpha\, I(t) - \gamma\, I(t)} \qquad \mathbf{(2)}$$

$$\mathbf{R'(t) = \alpha\, I(t)} \qquad \mathbf{(3)}$$

$$\mathbf{D'(t) = \gamma\, I(t)} \qquad \mathbf{(4)}$$

This model depends on three parameters: α (recovery rate per unit of time), β (infected rate per unit of time), and γ (death rate per unit of time).

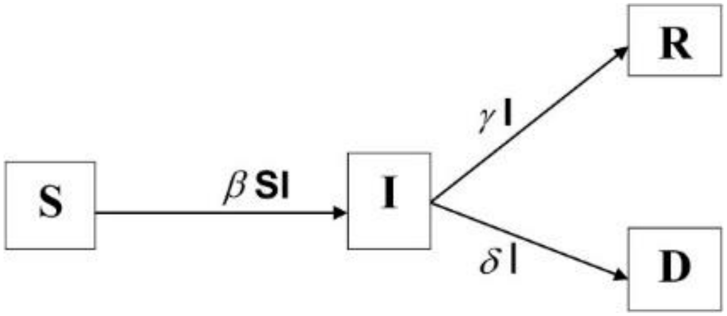

Figure 1: SIRD Model

For apply SIRD Model on graph, the following code is used:

```
def sird_model(y, x, beta, gamma, delta):
    S = -beta * y[0] * y[1] / N
    R = gamma * y[1]
    D = delta * y[1]
    I = - (S + R + D)
    return S, I, R, D
```

CHAPTER 3

LITERATURE REVIEW

3.1 Introduction

In this chapter the literature review for “Applying Statistical inference techniques to smoothen COVID19 data and predict future infection” has been discussed. In this section, some similar paper summary has been written.

Governments can make the best judgments possible with precise data to keep people safe during pandemics like the COVID-19 coronavirus. Data dependability is critical in such situations, therefore outlier detection is a critical, if not unavoidable, issue. Outliers are frequently regarded as the most intriguing observations, as their differences from the data majority may lead to important findings in the field. Outlier identification has also been studied in the context of multivariate functional data, which consists of smooth functions of various parameters, which are frequently generated from measurements taken at separate times. Only relative information in terms of log-ratios between these parts is considered significant for the analysis since the underlying data are treated as compositions, with the compositional portions providing the multivariate information. Using this relative knowledge, the multivariate functional data must be deduced as smooth functions. Following that, proven multivariate functional outlier detection algorithms can be employed, but the functional data must be presented in an acceptable space for interpretation. The concept is demonstrated using publicly available data from the COVID-19 pandemic to identify nations with unusual trends. [1]

Because of its rapid global spread, the coronavirus outbreak (COVID-19) poses a public health threat. Because of its extraordinary speed and reach, various countries have implemented a variety of remedies, including lockdowns, travel restrictions, and social distancing mandates. Understanding the latent dynamics of the disease's progression and the effectiveness of intervention programs is critical for controlling and preventing the spread of COVID-19. Hidden Markov models (HMMs) account for both randomness and uncertainty in spatiotemporal dynamics. In this article, they employ an overall HMM to infer the severity state on tiny geographical states or regions in the United States and Italy as test cases, using COVID-19 data from many countries, including the United States, several European countries, and countries with strong control laws. They also visualize the time evolution and dissemination across regions by aggregating the severity level of each location over a fixed time period. Such an analysis and visualization provide calibrated suggestions for interventions and responses. The results of HMM modeling is congruent with what has been observed in Italy and the United States, and these models can be used by policymakers as visualization and proactive decision support tools. [2]

On the basis of the epiMOX dashboard, which deals with data on epidemic trends and outbreaks in Italy from late February 2020, an analysis of the COVID-19 epidemic is provided. By fostering a deeper interpretation of accessible data through numerous crucial epidemic indicators, their analysis enables a quick awareness of the previous epidemic evolution as well as current tendencies. In addition, they supplement the epiMOX dashboard with SUIHTER, a predictive tool based on an epidemiological compartmental model for forecasting epidemic progression in the near future. [3]

During the coronavirus disease 2019 (COVID-19) pandemic, social distancing techniques such as school and public place closures, physical separation, and cancellation of large gatherings1 were undertaken to slow the spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Although children have been spared from severe SARS-CoV-2 infections that necessitate hospitalization,2,3 it is unclear if social distancing strategies are linked to pediatric inpatient use for other respiratory viruses. They investigated the impact of the COVID-19 pandemic and social distancing measures on pediatric hospitalizations at a freestanding children's hospital to fill this knowledge gap. [4]

The coronavirus disease pandemic of 2019 (COVID-19) has claimed the lives of millions of people in 216 countries throughout the world. Scientists are working around the world to better understand the nature of this devastating virus and develop a possible vaccine. This research has been presented in the best possible light as a contribution to the study of this fatal virus in its first and second waves. This research proposes a novel way for comparing devastating viral propagation in two waves throughout five of the world's most infected countries, as well as the virus's death rate, in order to gain a clear picture of the disease's behavior. In the first and second waves of this study, the number of deaths per day and the number of infected cases per day of the most impacted countries, the United States, Brazil, Russia, India, and the United Kingdom, were studied. For the COVID-19 mandated data sets, the correlation fractal dimension was estimated, and the death rate was compared using the correlation fractal dimension estimate curve. The statistical tool analysis of variance was also employed to support the proposed method's performance. [5]

The goal of this study is to contribute to a better understanding of the COVID-19 outbreak in Italy. To that purpose, we created a modified Susceptible-Infected-Recovered-Deceased (SIRD) model for the contagion, and we used official pandemic data to define the model's parameters. There are two primary non-standard components to our approach. The first is that model parameters can be time-varying, allowing us to reflect possible changes in epidemic behavior, such as those caused by government-imposed containment measures, changes in epidemic characteristics, or the impact of advanced antiviral medicines. The time-varying parameters are expressed as linear combinations of basic functions, and sparse identification techniques are used to infer them from data. The second non-standard angle dwells in the way that we consider as model boundaries additionally the underlying number of vulnerable people, as well as the proportionality factor relating the distinguished number of up-sides with the real (and obscure) number of tainted people. Recognizing the model boundaries adds up to a non-curved ID issue that we settle through a settled methodology, comprising in a one-layered matrix search in the external circle. [7]

Covid-19 is a highly contagious virus that has nearly paralyzed the global economy. Its ability to transmit human-to-human and surface-to-human information plunges the planet into chaos. The goal of this research is to predict the future conditions of a novel Coronavirus in order to reduce its impact. In India and the United States, we suggested a deep learning-based comparative analysis of Covid-19 instances. The datasets of Covid-19 confirmed and mortality cases are taken into account. To create the proposed approach and forecast the Covid-19 instances for one month ahead, recurrent neural network (RNN) based variations of long short term memory (LSTM) such as Stacked LSTM, Bi-directional LSTM, and Convolutional LSTM are employed. For all four datasets from both nations, Convolution LSTM beat the other two models, predicting Covid-19 cases with high accuracy and very low error. The upward/downward trend of anticipated Covid-19 cases is also graphically represented, which can help researchers and policymakers reduce mortality and morbidity rates by streaming the Covid-19 in the proper direction. [9]

CHAPTER 4

RESULT AND DISCUSSION

4.1 Introduction

This chapter discusses the results of the outcome. In section 4.2, the expected output for daily infection and deaths for last one year has been discussed. In section 4.3 – 4.9, the expected output for Identification of Outliers on Daily Infection and Deaths for Last One Year has been discussed. In section 4.4, the expected output for Smoothing Curves on data of last one year after applying Savitzky-Golay (Savgol) filter has been discussed. In section 4.5, the expected output for SIRD Model Fitted Curves for Last One Year has been discussed.

4.2 Daily Infection and Deaths for Last One Year

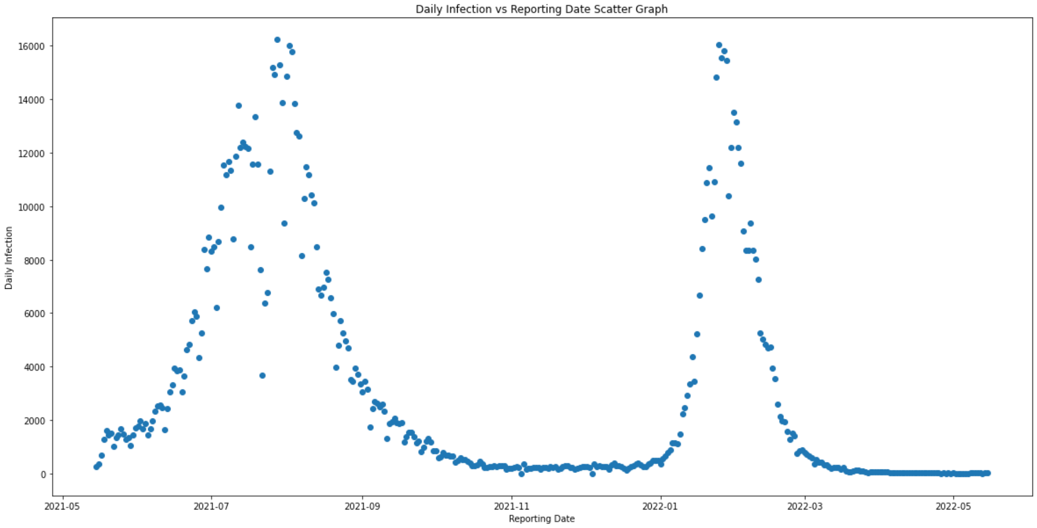

Figure 2: Daily Infection vs Reporting Date

This is the scatter graph of Daily Infection vs Reporting Date. Here, the graph is plotted on one year of data from 15th May 2021 to 15th May 2022.

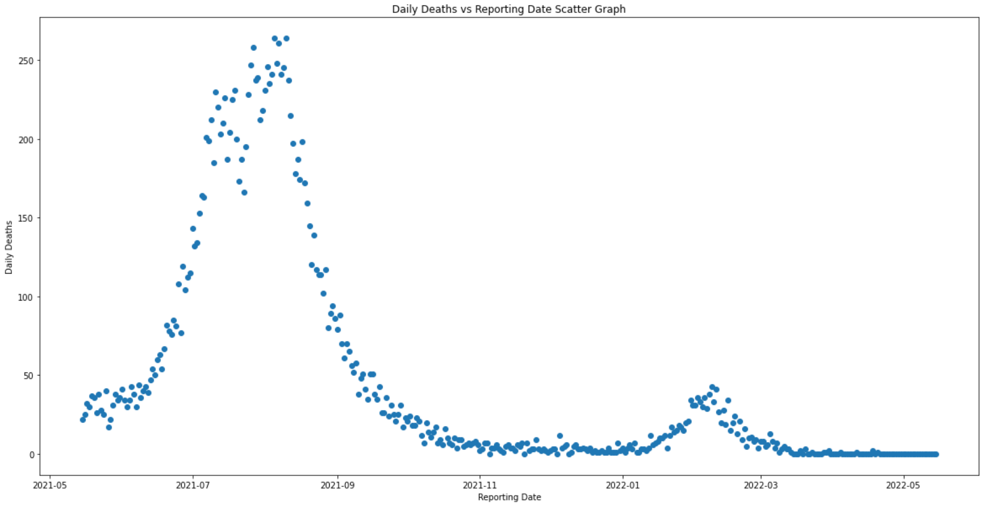

Figure 3: Daily Deaths vs Reporting Date

This is the scatter graph of Daily Deaths vs Reporting Date. Here, the graph is plotted on one year of data from 15th May 2021 to 15th May 2022.

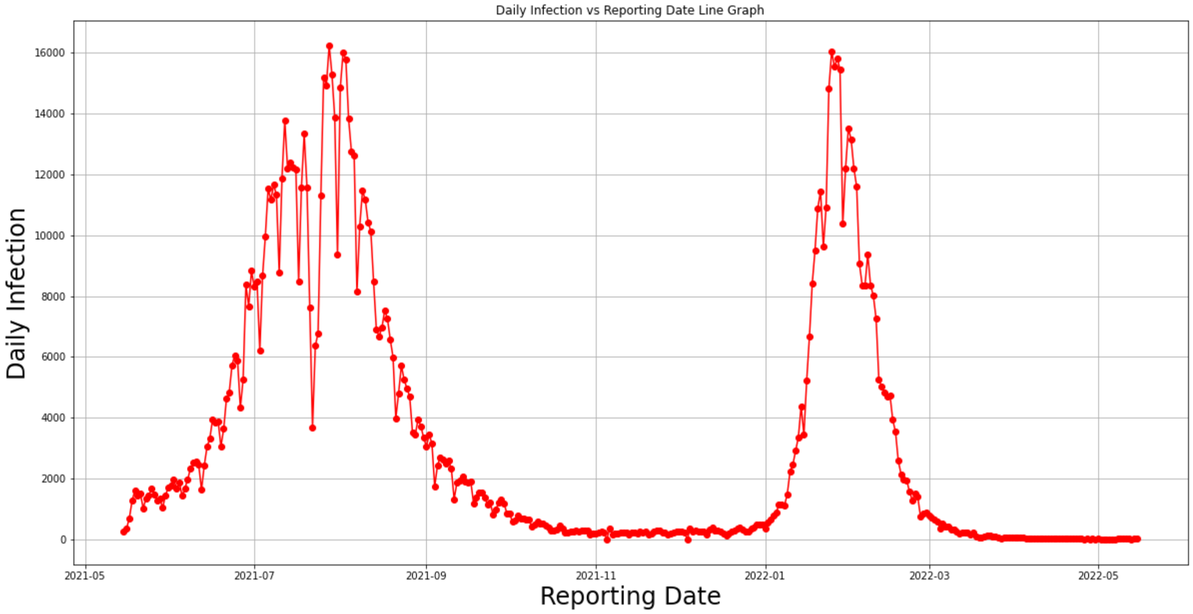

Figure 4: Daily Infection vs Reporting Date

This is the line graph of Daily Infection vs Reporting Date. Here, the graph is plotted on one year of data from 15th May 2021 to 15th May 2022.

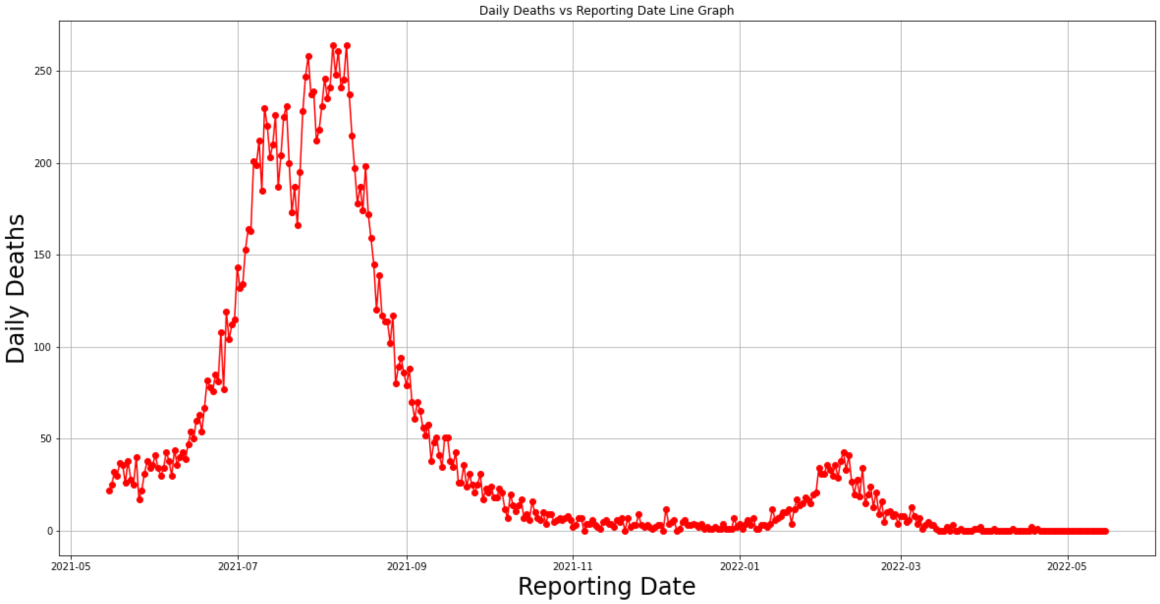

Figure 5: Daily Deaths vs Reporting Date

This is the line graph of Daily Deaths vs Reporting Date. Here, the graph is plotted on one year of data from 15th May 2021 to 15th May 2022.

4.3 Identification of Outliers on Daily Infection from 15th May 2021 to 11th September 2021

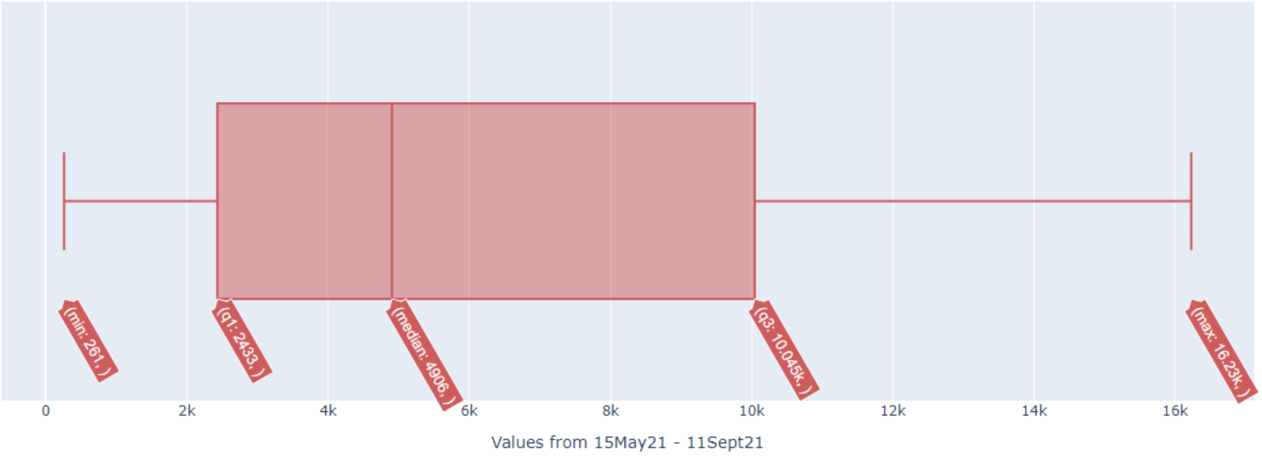

Figure 6: Boxplot of Daily Infection from 15th May 2021 to 11th September 2021

This is a boxplot of daily infection data from 15th May 2021 to 11th September 2021. Here, there is outlier.

4.4 Identification of Outliers on Daily Infection from 12th September 2021 to 9th January 2022

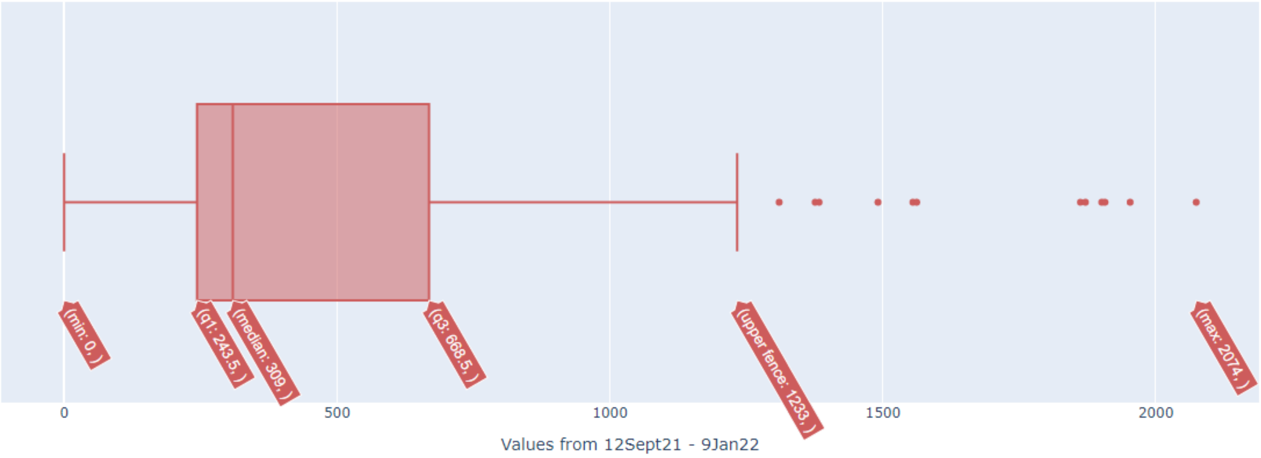

Figure 7: Boxplot of Daily Infection from 12th September 2021 to 9th January 2022

This is a boxplot of daily infection data from 12th September 2021 to 9th January 2022. Here, there are 12 outliers.

Reporting Date	Value
2021-09-12	1871
2021-09-13	1953
2021-09-14	2074
2021-09-15	1901
2021-09-16	1862
2021-09-17	1907
2021-09-19	1383
2021-09-20	1555
2021-09-21	1562
2021-09-22	1376
2021-09-28	1310
2022-01-09	1491

Table 1: Table of Outlier Values of Daily Infection from 12th September 2021 to 9th January 2022

This is the table of 12 outlier values and the corresponding reporting date of daily infection data from 12th September 2021 to 9th January 2022.

4.5 Identification of Outliers on Daily Infection from 10th January 2022 to 15th May 2022

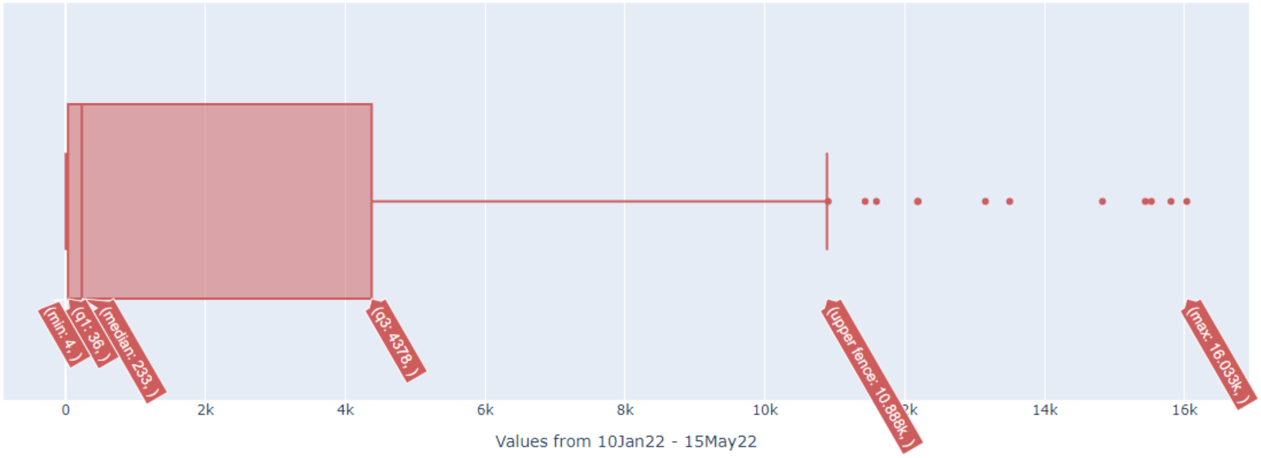

Figure 8: Boxplot of Daily Infection from 10th January 2022 to 15th May 2022

This is a boxplot of daily infection data from 10th January 2021 to 15th May 2022. Here, there are 12 outliers.

Reporting Date	Value
2022-01-21	11434
2022-01-23	10906
2022-01-24	14828
2022-01-25	16033
2022-01-26	15527
2022-01-27	15807
2022-01-28	15440
2022-01-30	12183
2022-01-31	13501
2022-02-01	13154
2022-02-02	12193
2022-02-03	11596

Table 2: Table of Outlier Values of Daily Infection from 10th January 2022 to 15th May 2022

This is the table of 12 outlier values and the corresponding reporting date of daily infection data from 10th January 2022 to 15th May 2022.

4.6 Identification of Outliers on Daily Deaths from 15th May 2021 to 11th September 2021

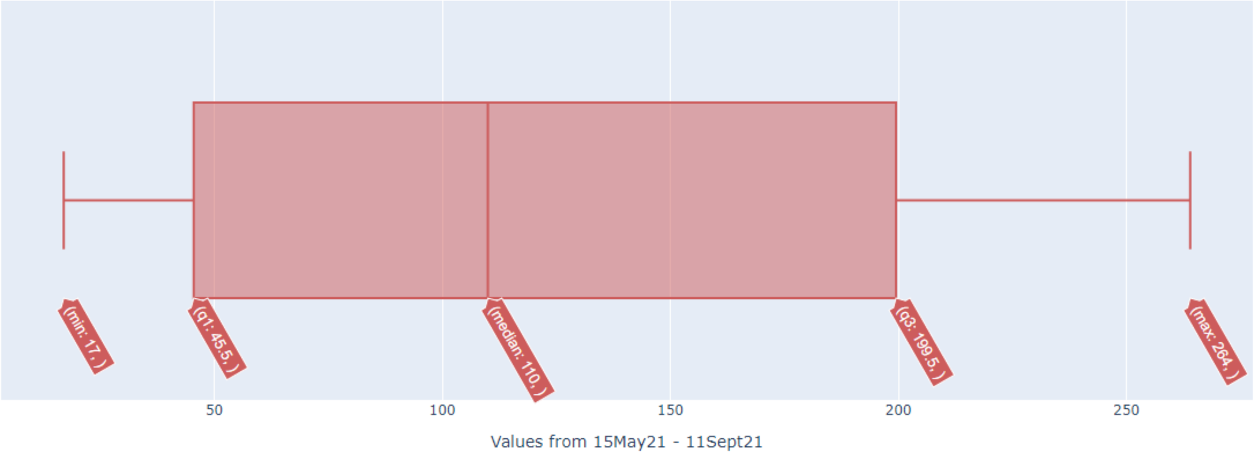

Figure 9: Boxplot of Daily Deaths from 15th May 2021 to 11th September 2021

This is a boxplot of daily deaths data from 15th May 2021 to 11th September 2021. Here, there is outlier.

4.7 Identification of Outliers on Daily Deaths from 12th September 2021 to 9th January 2022

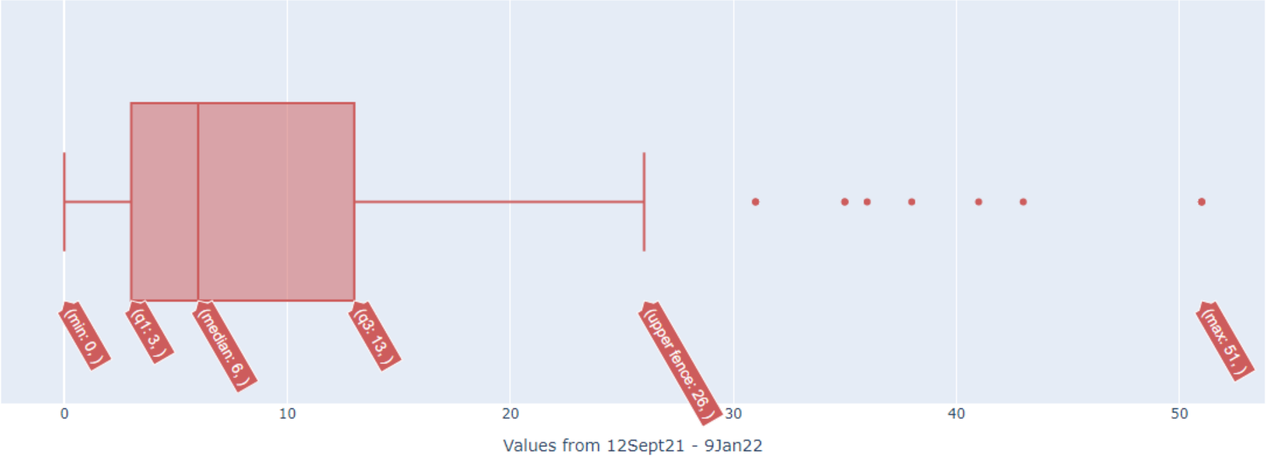

Figure 10: Boxplot of Daily Deaths from 12th September 2021 to 9th January 2022

This is a boxplot of daily deaths data from 12th September 2021 to 9th January 2022. Here, there are 11 outliers.

Reporting Date	Value
2021-09-12	51
2021-09-13	41
2021-09-14	35
2021-09-15	51
2021-09-16	51
2021-09-17	38
2021-09-18	35
2021-09-19	43
2021-09-22	36
2021-09-24	31
2021-09-28	31

Table 3: Table of Outlier Values of Daily Deaths from 12th September 2021 to 9th January 2022

This is the table of 11 outlier values and the corresponding reporting date of daily deaths data from 12th September 2021 to 9th January 2022.

4.8 Identification of Outliers on Daily Deaths from 10th January 2022 to 15th May 2022

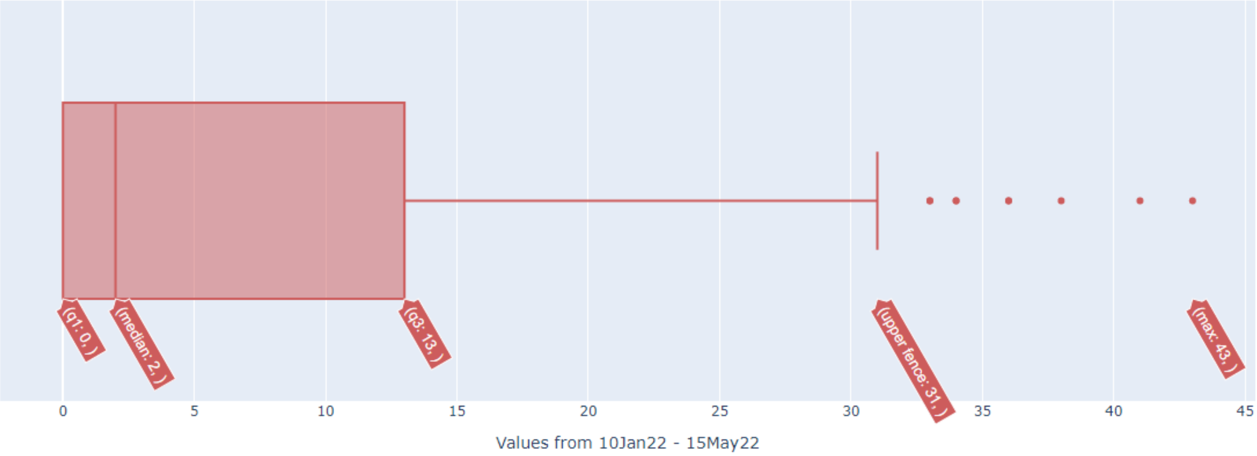

Figure 11: Boxplot of Daily Deaths from 10th January 2022 to 15th May 2022

This is a boxplot of daily deaths data from 10th January 2021 to 15th May 2022. Here, there are 9 outliers.

Reporting Date	Value
2022-01-30	34
2022-02-02	36
2022-02-03	33
2022-02-05	36
2022-02-07	38
2022-02-08	43
2022-02-09	33
2022-02-10	41
2022-02-15	34

Table 4: Table of Outlier Values of Daily Deaths from 10th January 2022 to 15th May 2022

This is the table of 9 outlier values and the corresponding reporting date of daily deaths data from 10th January 2022 to 15th May 2022.

4.9 Smoothing Curves on Daily Infection after applying Savitzky-Golay Filter with Polynomial Order 3

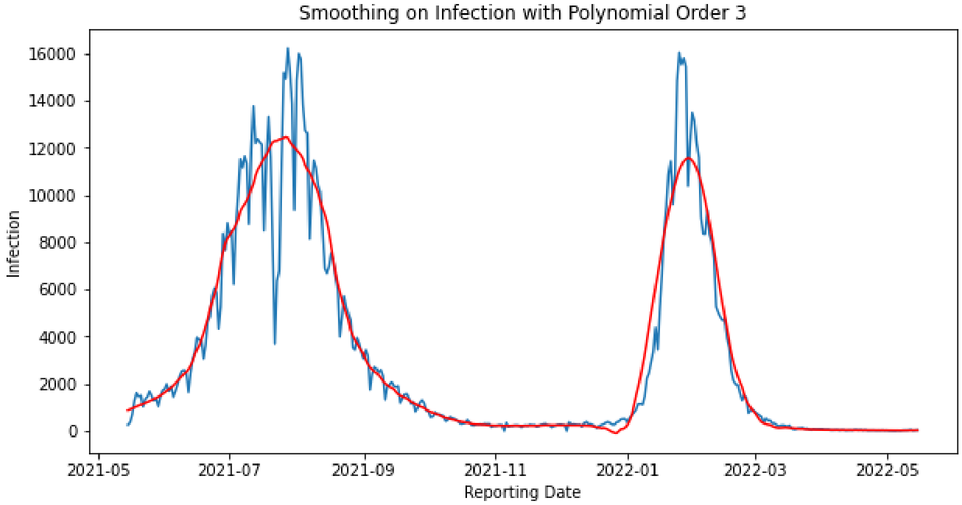

Figure 12: Daily Infection vs Reporting Date Graph (Blue Line) and Smoothing Curve with Polynomial Order 3 (Red Line)

Here, in this graph, the Reporting Date is in the x-axis and the value of Daily Infection is in the y-axis. The blue line is Daily Infection vs Reporting Date graph. The red line is the smoothing graph which we have done with applying Savgol-Filter with polynomial order 3.

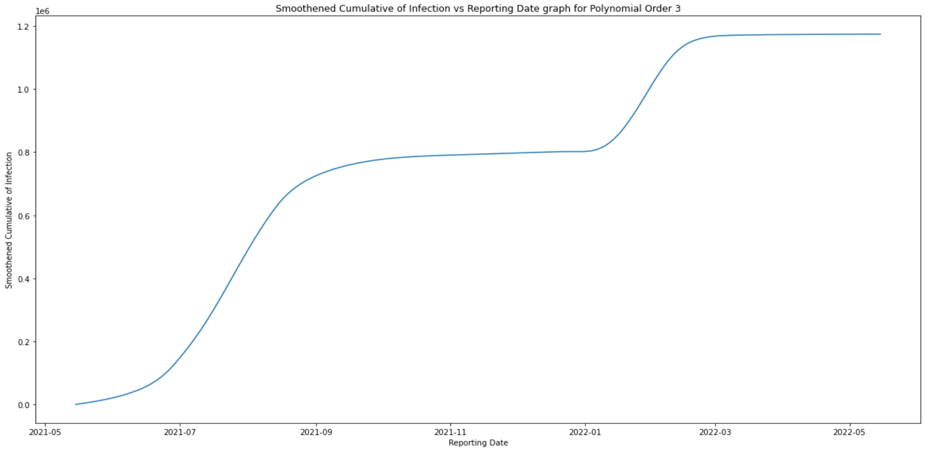

Figure 13: Smoothened Cumulative of Infection vs Reporting Date Graph for Polynomial Order 3

This is the smoothened cumulative of Infection vs Reporting Date Graph for polynomial order 3. First, we find out the smoothened values of Infected from the smoothened curve of Infected and then we calculate the smoothened cumulative of Infection and plot this graph.

4.10 Smoothing Curves on Daily Infection after applying Savitzky-Golay Filter with Polynomial Order 4

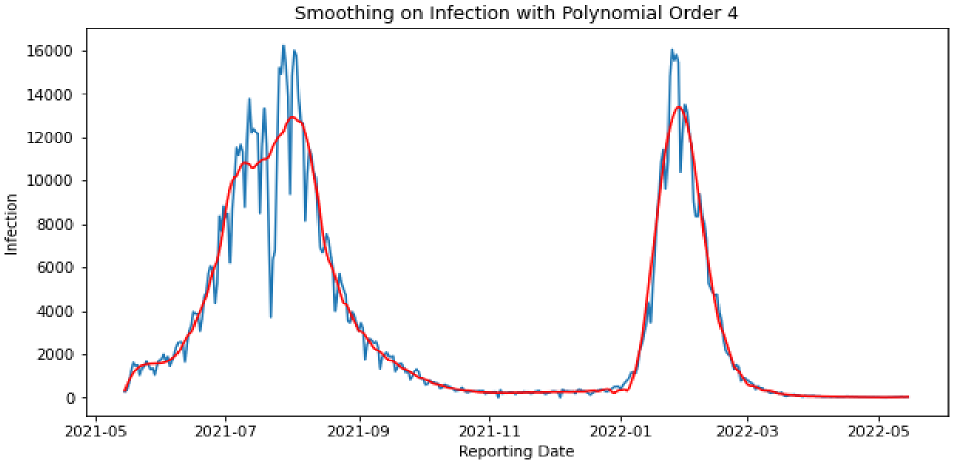

Figure 14: Daily Infection vs Reporting Date Graph (Blue Line) and Smoothing Curve with Polynomial Order 4 (Red Line)

Here, in this graph, the Reporting Date is in the x-axis and the value of Daily Infection is in the y-axis. The blue line is Daily Infection vs Reporting Date graph. The red line is the smoothing graph which we have done with applying Savgol-Filter with polynomial order 4.

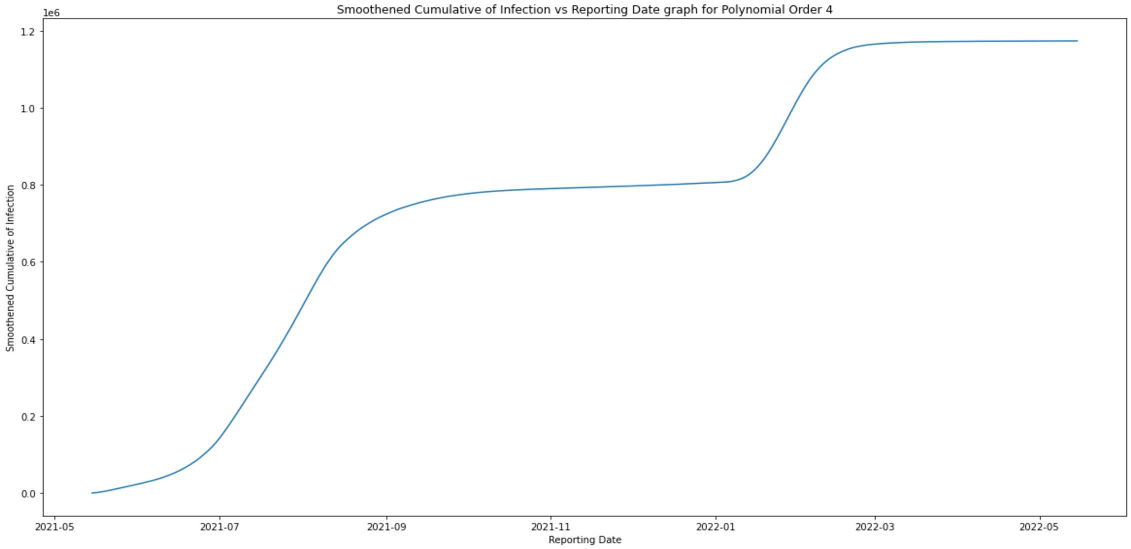

Figure 15: Smoothened Cumulative of Infection vs Reporting Date Graph for Polynomial Order 4

This is the smoothened cumulative of Infection vs Reporting Date Graph for polynomial order 4. First, we find out the smoothened values of Infected from the smoothened curve of Infected and then we calculate the smoothened cumulative of Infection and plot this graph.

4.11 Smoothing Curves on Daily Infection after applying Savitzky-Golay Filter with Polynomial Order 5

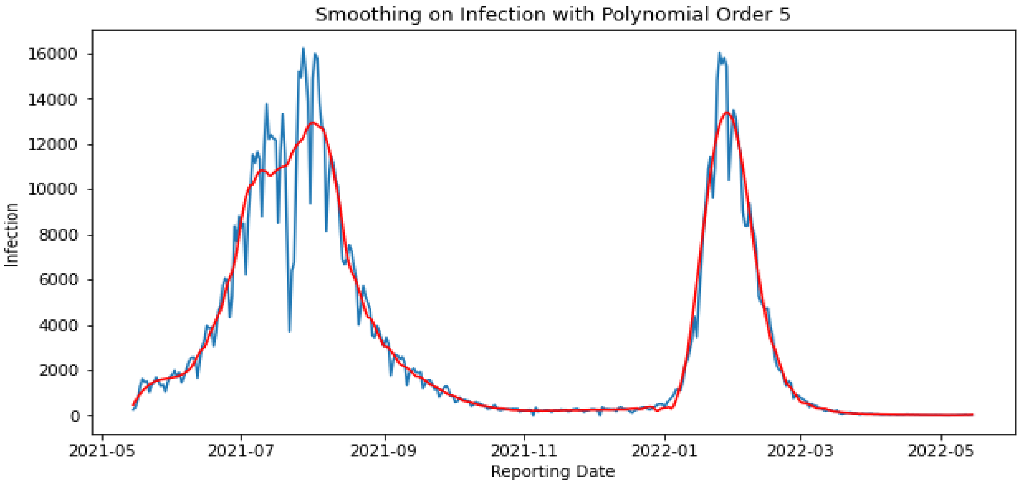

Figure 16: Daily Infection vs Reporting Date Graph (Blue Line) and Smoothing Curve with Polynomial Order 5 (Red Line)

Here, in this graph, the Reporting Date is in the x-axis and the value of Daily Infection is in the y-axis. The blue line is Daily Infection vs Reporting Date graph. The red line is the smoothing graph which we have done with applying Savgol-Filter with polynomial order 5.

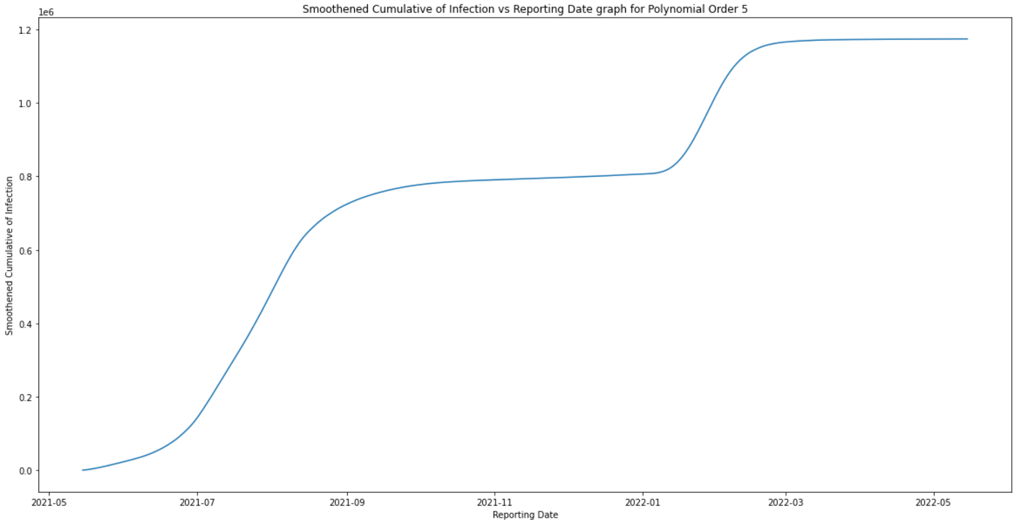

Figure 17: Smoothened Cumulative of Infection vs Reporting Date Graph for Polynomial Order 5

This is the smoothened cumulative of Infection vs Reporting Date Graph for polynomial order 5. First, we find out the smoothened values of Infected from the smoothened curve of Infected and then we calculate the smoothened cumulative of Infection and plot this graph.

4.12 Smoothing Curves on Daily Deaths after applying Savitzky-Golay Filter with Polynomial Order 3

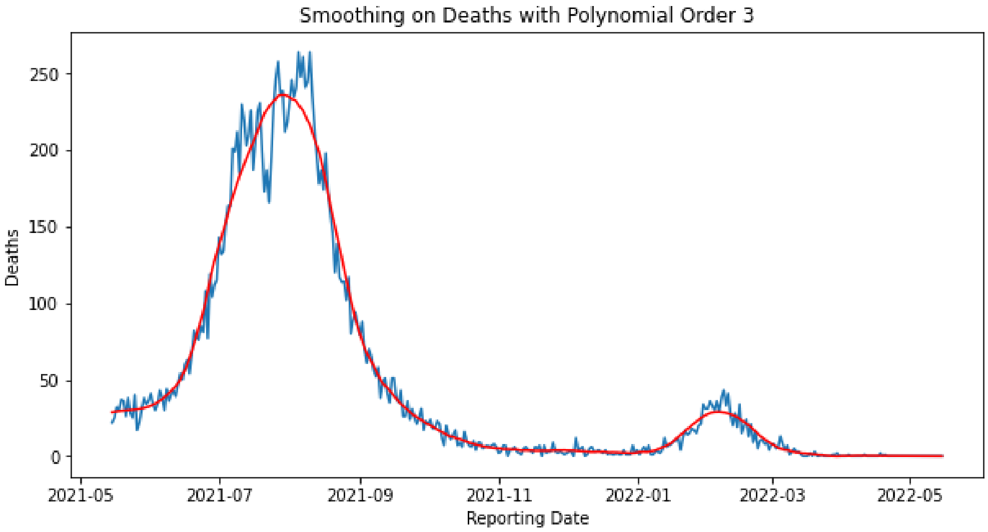

Figure 18: Daily Deaths vs Reporting Date Graph (Blue Line) and Smoothing Curve with Polynomial Order 3 (Red Line)

Here, in this graph, the Reporting Date is in the x-axis and the value of Daily Deaths is in the y-axis. The blue line is Daily Deaths vs Reporting Date graph. The red line is the smoothing graph which we have done with applying Savgol-Filter with polynomial order 3.

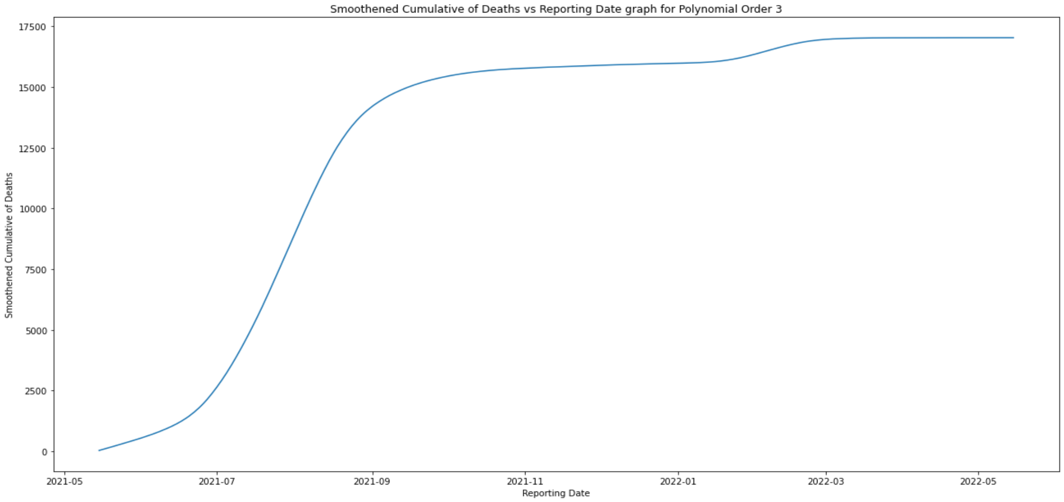

Figure 19: Smoothened Cumulative of Deaths vs Reporting Date Graph for Polynomial Order 3

This is the smoothened cumulative of Deaths vs Reporting Date Graph for polynomial order 3. First, we find out the smoothened values of Deaths from the smoothened curve of Infected and then we calculate the smoothened cumulative of Deaths and plot this graph.

4.13 Smoothing Curves on Daily Deaths after applying Savitzky-Golay Filter with Polynomial Order 4

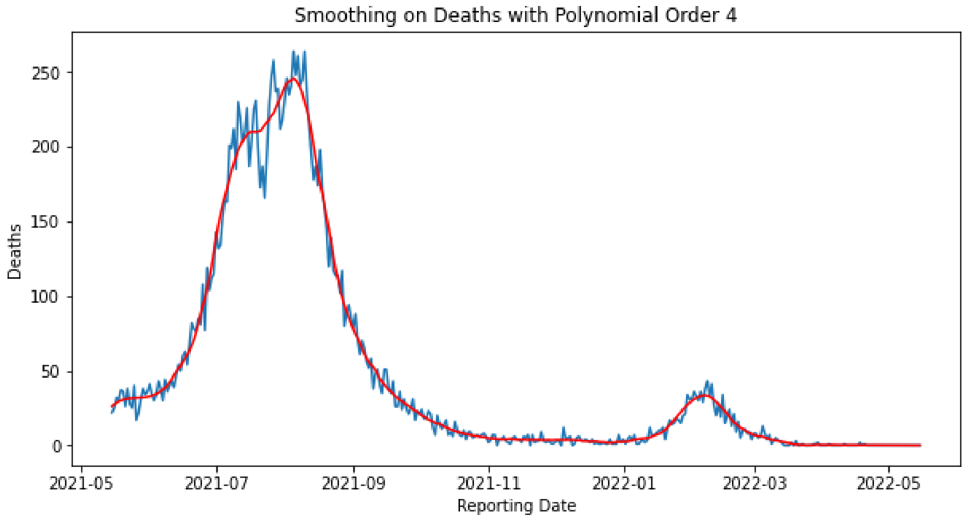

Figure 20: Daily Deaths vs Reporting Date Graph (Blue Line) and Smoothing Curve with Polynomial Order 4 (Red Line)

Here, in this graph, the Reporting Date is in the x-axis and the value of Daily Deaths is in the y-axis. The blue line is Daily Deaths vs Reporting Date graph. The red line is the smoothing graph which we have done with applying Savgol-Filter with polynomial order 4.

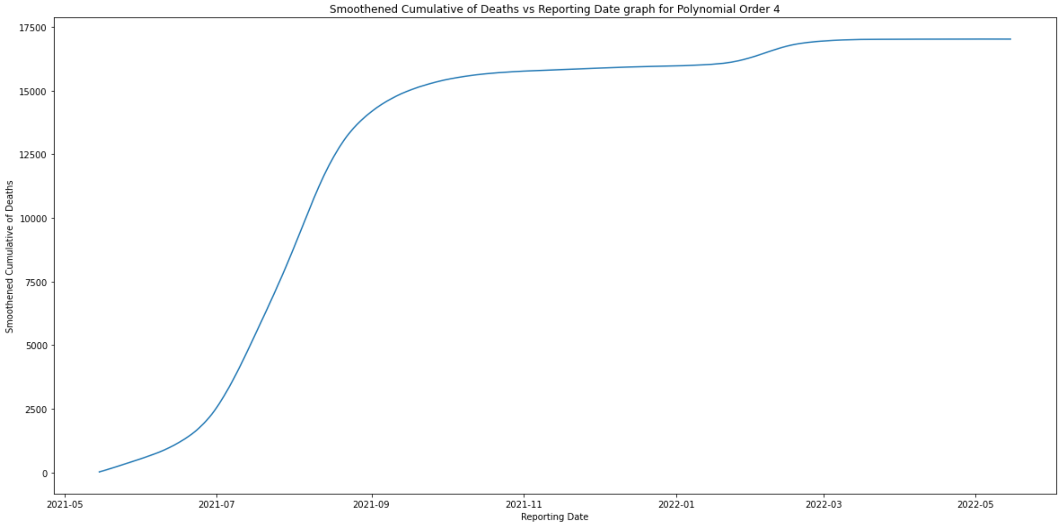

Figure 21: Smoothened Cumulative of Deaths vs Reporting Date Graph for Polynomial Order 4

This is the smoothened cumulative of Deaths vs Reporting Date Graph for polynomial order 4. First, we find out the smoothened values of Deaths from the smoothened curve of Infected and then we calculate the smoothened cumulative of Deaths and plot this graph.

4.14 Smoothing Curves on Daily Deaths after applying Savitzky-Golay Filter with Polynomial Order 5

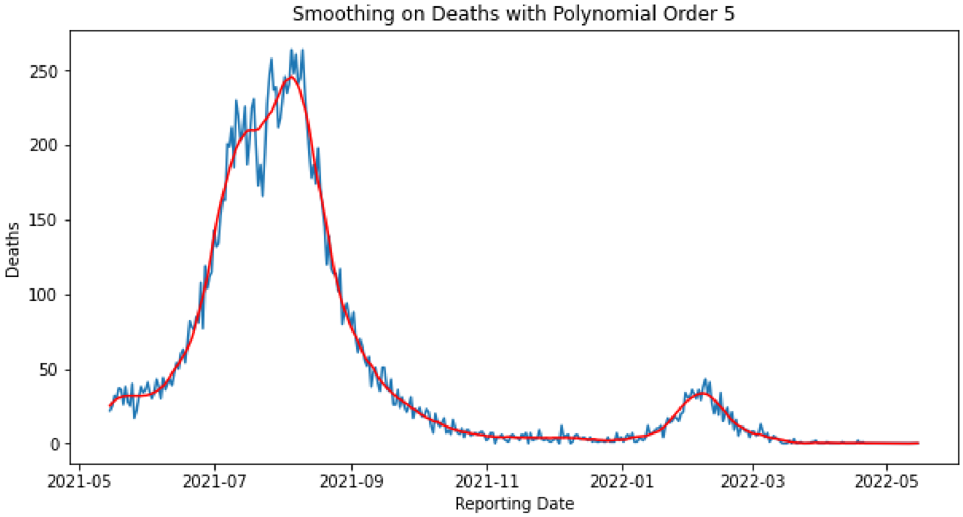

Figure 22: Daily Deaths vs Reporting Date Graph (Blue Line) and Smoothing Curve with Polynomial Order 5 (Red Line)

Here, in this graph, the Reporting Date is in the x-axis and the value of Daily Deaths is in the y-axis. The blue line is Daily Deaths vs Reporting Date graph. The red line is the smoothing graph which we have done with applying Savgol-Filter with polynomial order 5.

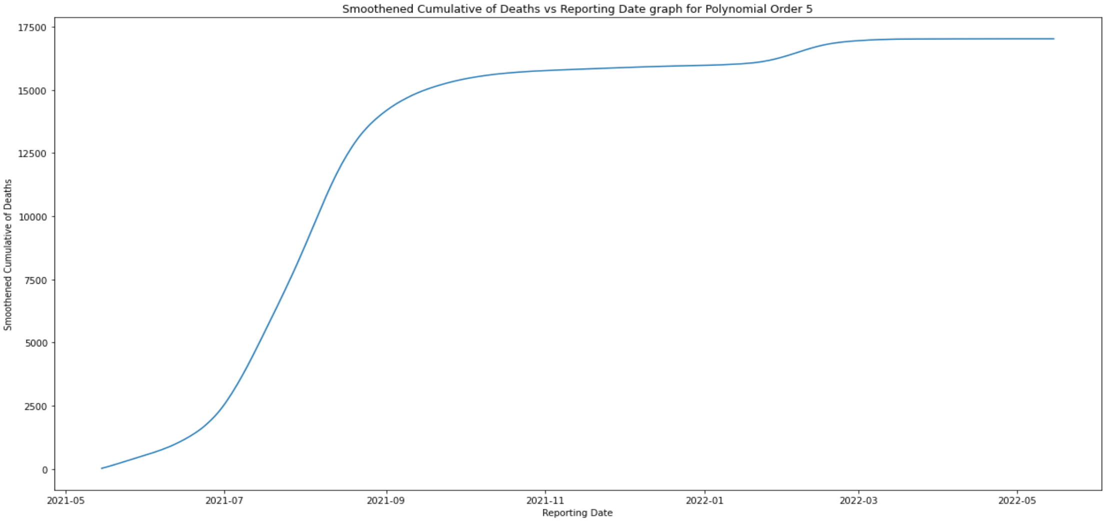

Figure 23: Smoothened Cumulative of Deaths vs Reporting Date Graph for Polynomial Order 5

This is the smoothened cumulative of Deaths vs Reporting Date Graph for polynomial order 5. First, we find out the smoothened values of Deaths from the smoothened curve of Infected and then we calculate the smoothened cumulative of Deaths and plot this graph.

4.15 SIRD Model Fitted Curves to Smoothened Cumulative and Infected (Cumulative) for Polynomial Order 3

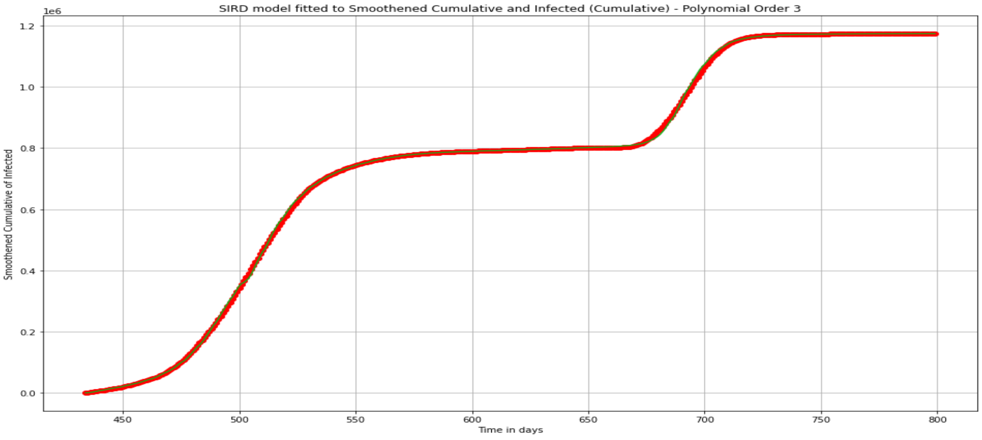

Figure 24: SIRD Model fitted to Smoothened Cumulative and Infected (Cumulative) – Polynomial Order 3

Here, in this graph, the Days since the first case (Time in days) is in the x-axis and the smoothened cumulative of Infected is in the y-axis. The red line is the smoothened cumulative of Infected vs Time in days graph for polynomial order 3. The green line is the Infected (Cumulative) vs Time in days graph.

4.16 SIRD Model Fitted Curves to Smoothened Cumulative and Infected (Cumulative) for Polynomial Order 4

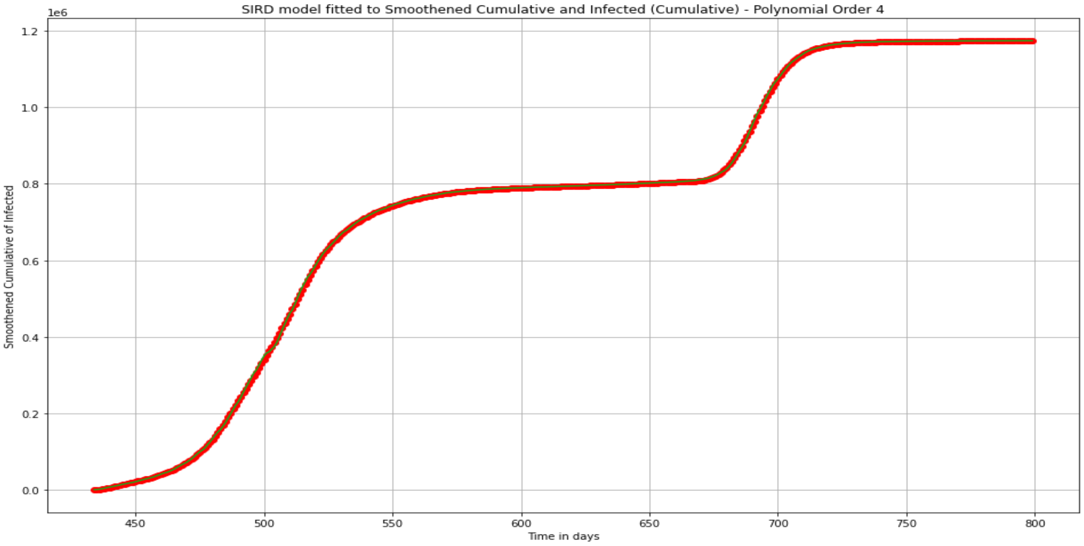

Figure 25: SIRD Model fitted to Smoothened Cumulative and Infected (Cumulative) – Polynomial Order 4

Here, in this graph, the Days since the first case (Time in days) is in the x-axis and the smoothened cumulative of Infected is in the y-axis. The red line is the smoothened cumulative of Infected vs Time in days graph for polynomial order 4. The green line is the Infected (Cumulative) vs Time in days graph.

4.17 SIRD Model Fitted Curves to Smoothened Cumulative and Infected (Cumulative) for Polynomial Order 5

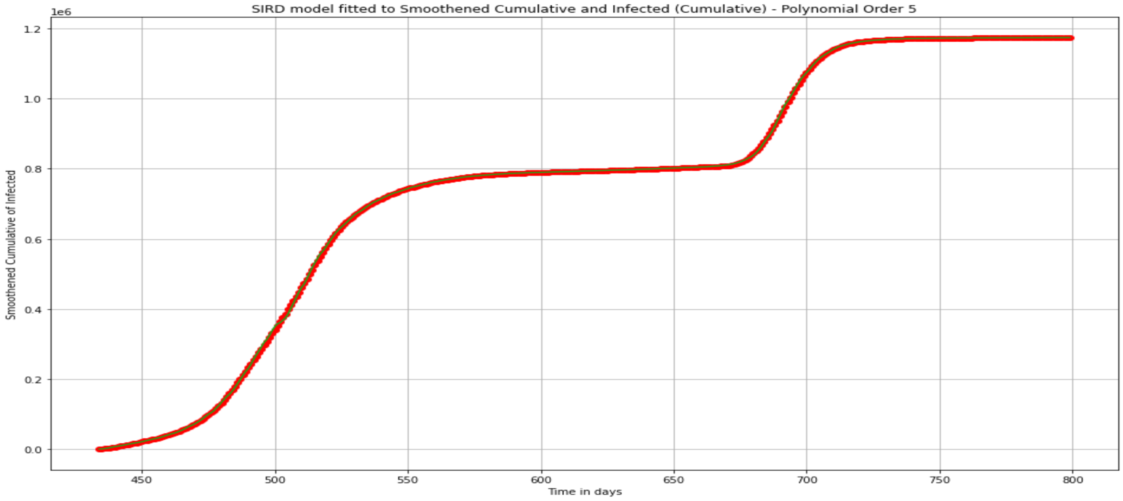

Figure 26: SIRD Model fitted to Smoothened Cumulative and Infected (Cumulative) – Polynomial Order 5

Here, in this graph, the Days since the first case (Time in days) is in the x-axis and the smoothened cumulative of Infected is in the y-axis. The red line is the smoothened cumulative of Infected vs Time in days graph for polynomial order 5. The green line is the Infected (Cumulative) vs Time in days graph.

4.18 SIRD Model Fitted Curves to Smoothened Cumulative and Deaths (Cumulative) for Polynomial Order 3

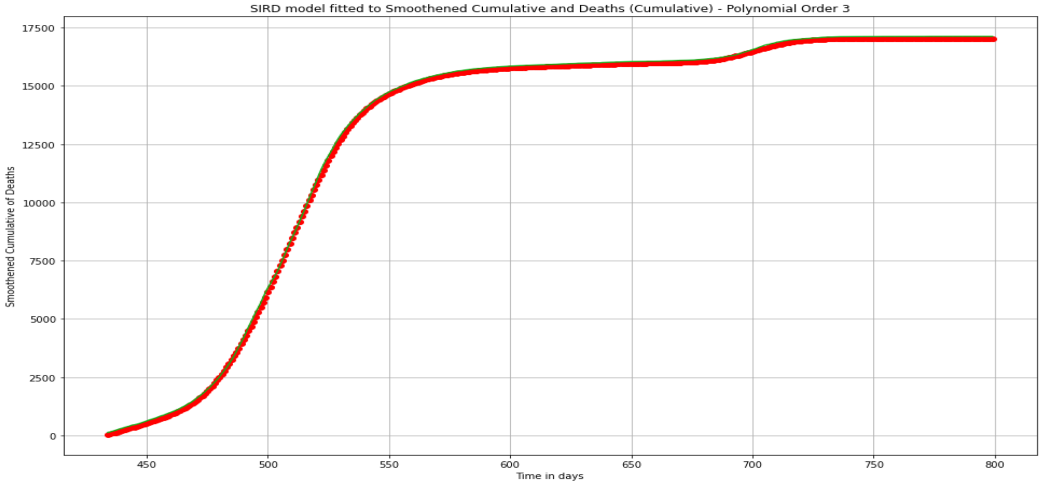

Figure 27: SIRD Model fitted to Smoothened Cumulative and Death (Cumulative) – Polynomial Order 3

Here, in this graph, the Days since the first case (Time in days) is in the x-axis and the smoothened cumulative of Deaths is in the y-axis. The red line is the smoothened cumulative of Deaths vs Time in days graph for polynomial order 3. The green line is the Deaths (Cumulative) vs Time in days graph.

4.19 SIRD Model Fitted Curves to Smoothened Cumulative and Deaths (Cumulative) for Polynomial Order 4

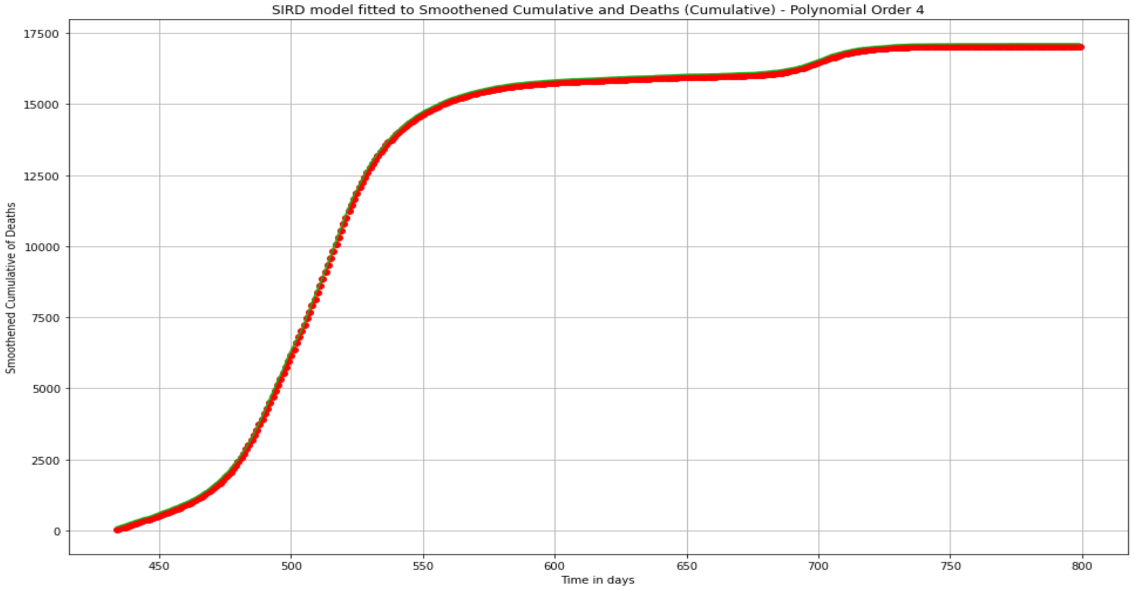

Figure 28: SIRD Model fitted to Smoothened Cumulative and Death (Cumulative) – Polynomial Order 4

Here, in this graph, the Days since the first case (Time in days) is in the x-axis and the smoothened cumulative of Deaths is in the y-axis. The red line is the smoothened cumulative of Deaths vs Time in days graph for polynomial order 4. The green line is the Deaths (Cumulative) vs Time in days graph.

4.20 SIRD Model Fitted Curves to Smoothened Cumulative and Deaths (Cumulative) for Polynomial Order 5

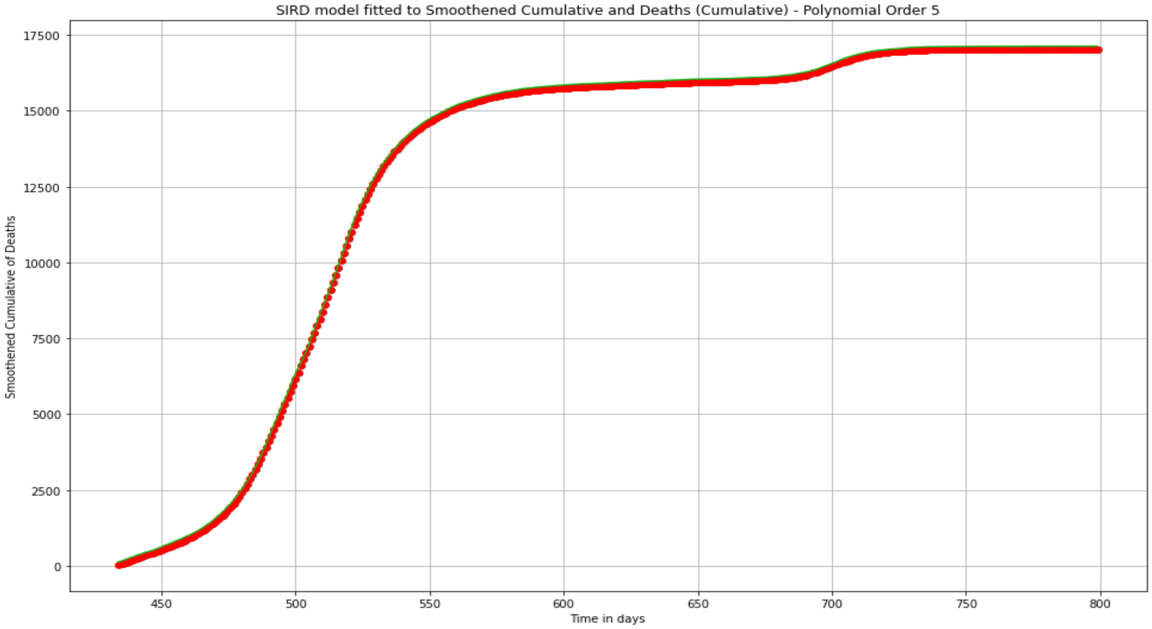

Figure 29: SIRD Model fitted to Smoothened Cumulative and Death (Cumulative) – Polynomial Order 5

Here, in this graph, the Days since the first case (Time in days) is in the x-axis and the smoothened cumulative of Deaths is in the y-axis. The red line is the smoothened cumulative of Deaths vs Time in days graph for polynomial order 5. The green line is the Deaths (Cumulative) vs Time in days graph.

4.21 Applying SIRD Model on Smoothened Cumulative of Infection - Polynomial Order 4 (Zoom of 50 Days from 450 – 500)

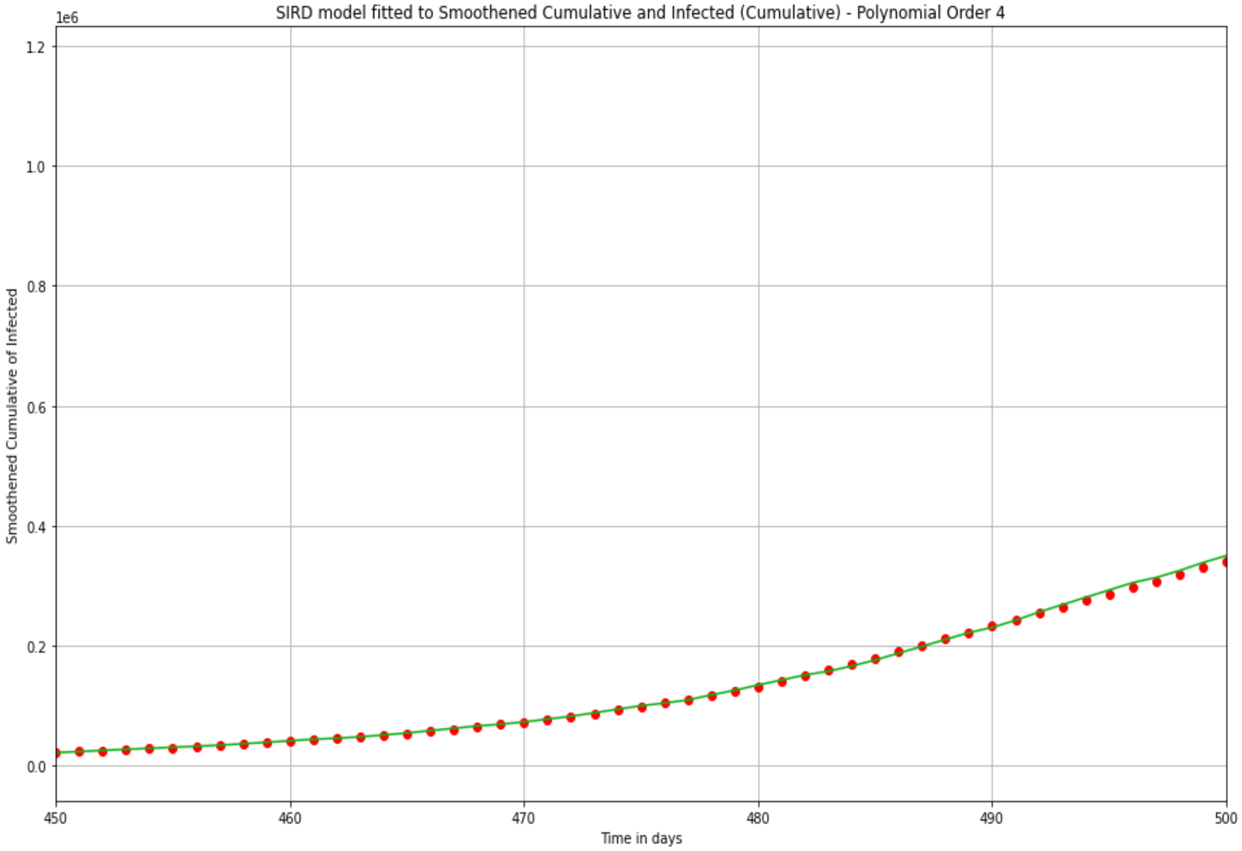

Figure 30: Applying SIRD Model on Smoothened Cumulative of Infection - Polynomial Order 4 (Zoom of 50 Days from 450 – 500)

Here, in this graph, the Days since the first case (Time in days) is in the x-axis and the smoothened cumulative of infection is in the y-axis. The red dotted line is the smoothened cumulative of Infection vs Time in days graph for polynomial order 4. The green line is the Infected (Cumulative) vs Time in days graph. The graph is zoomed to 450 to 500 days which is 50 days in total. After zooming, we can see a slight changes in the graph

CHAPTER 5

CONCLUSION

This report points to appear the show condition of COVID19 data of Infection and Deaths (Daily and Cumulative). There are some errors in the data for different variation such as human error and to find out the errors in the data by applying mathematical methods and statistical inference techniques to smoothen COVID19 data and predict future infection. It lets us to explore and analyze data using appealing graphics and simple, yet innovative methodologies. It can assist users in staying informed and up to date on the spread of COVID19 and related issues at all times. I have collected the COVID19 data for last one year. I have stored the data in an excel file of extension .xlsx. Then, I have visualized the graph by plotting Infection/Deaths in y-axis and Reporting Date in x-axis. To find out the abnormal values, I found out the outliers of daily infection and deaths from the data by generating boxplot. As the abnormal values increase the error of statistical data which can cause bias or influence, we have to correct it. I have applied savgol-filter to smooth the graph of Infection and Deaths in different polynomial order which is 3,4 and 5. After plotting graphs in different polynomial order and window size, we have seen some changes in the graphs. The graph is being stretched a little bit on to the top. At last, I have applied standard SIRD Model to analyze the number of susceptible, infected, recovered, and deceased people from the COVID19 dataset. I have zoomed in the figure to a portion of one week/month to see the changes in the graphs. I did not used the other SIRD Models because they are rational model. We have done our code in python languages. As a result of our work, other people may be inspired to study and analysis on the data of COVID19.

CHAPTER 6

REFERENCES

[1] Rieser, Christopher & Filzmoser, Peter. (2022). Outlier Detection for Pandemic-Related Data Using Compositional Functional Data Analysis. 10.1007/978-3-030-78334-1_12. [Accessed May 19, 2022].

[2] Zhou, Shanglin & Braca, Paolo & Gaglione, Domenico & Millefiori, Leonardo & Marano, Stefano & Willett, Peter & Pattipati, Krishna. (2021). Application of Hidden Markov Models to Analyze, Group and Visualize Spatio-Temporal COVID-19 Data. IEEE Access. PP. 10.1109/ACCESS.2021.3114364. [Accessed May 19, 2022].

[3] N. Parolini, G. Ardenghi, L. Dede', and A. Quarteroni, "A mathematical dashboard for the analysis of Italian COVID-19 epidemic data," Int. j. numer. method. biomed. eng., vol. 37, no. 9, p. e3513, 2021. [Accessed May 19, 2022].

[4] J. L. Wilder, C. R. Parsons, A. S. Growdon, S. L. Toomey, and J. M. Mansbach, "Pediatric hospitalizations during the COVID-19 pandemic," Pediatrics, vol. 146, no. 6, p. e2020005983, 2020.

[5] D. Easwaramoorthy, A. Gowrisankar, A. Manimaran, S. Nandhini, L. Rondoni, and S. Banerjee, "An exploration of fractal-based prognostic model and comparative analysis for second wave of COVID-19 diffusion," *Nonlinear Dyn.*, vol. 106, no. 2, pp. 1375–1395, 2021. [Accessed May 19, 2022].

[6] Braca, P., Gaglione, D., Marano, S. *et al.* Decision support for the quickest detection of critical COVID-19 phases. *Sci Rep* **11,** 8558 (2021). https://doi.org/10.1038/s41598-021-86827-6

[7] Giuseppe C. Calafiore, Carlo Novara, Corrado Possieri, A time-varying SIRD model for the COVID-19 contagion in Italy, Annual Reviews in Control, Volume 50, 2020, Pages 361-372, ISSN 1367-5788, https://doi.org/10.1016/j.arcontrol.2020.10.005. [Accessed May 19, 2022].

[8] Sam, Shem. (2020). Exploring the Statistical Significance of Africa's COVID-19 Data. International Journal of Applied Mathematics and Statistics. 5. [Accessed May 19, 2022].

[9] Shastri, S. et al., 2020. Time series forecasting of covid-19 using Deep Learning Models: India-USA Comparative Case Study. *Chaos, solitons, and fractals*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7440083/ [Accessed May 19, 2022].

[10] A. Poddar and M. Poddar, "Covid-19 data visualization and data analytics with a smart standalone mobile application," in *2020 IEEE 17th India Council International Conference (INDICON)*, 2020, pp. 1–6.

[11] J. Fernández-Villaverde and C. I. Jones, "Estimating and simulating a SIRD model of COVID-19 for many countries, states, and cities," *J. Econ. Dyn. Control*, no. 104318, p. 104318, 2022.

[12] S. Shringi, H. Sharma, P. N. Rathie, J. C. Bansal, and A. Nagar, "Modified SIRD model for COVID-19 spread prediction for Northern and Southern states of India," *Chaos Solitons Fractals*, vol. 148, no. 111039, p. 111039, 2021.

[13] V. Martínez, "A modified SIRD model to study the evolution of the COVID-19 pandemic in Spain," *Symmetry (Basel)*, vol. 13, no. 4, p. 723, 2021.

[14] B. Basti, N. Hammami, I. Berrabah, F. Nouioua, R. Djemiat, and N. Benhamidouche, "Stability analysis and existence of solutions for a modified SIRD model of COVID-19 with fractional derivatives," *Symmetry (Basel)*, vol. 13, no. 8, p. 1431, 2021.

[15] J. Rubio-Herrero and Y. Wang, "A flexible rolling regression framework for time-varying SIRD models: Application to COVID-19," *arXiv [q-bio.PE]*, 2021.