SIR Model for Infectious diseases

First Degree Thesis 2016-2017

BITS Pilani KK Birla Goa Campus

Submitted in partial fulfillment of the requirements of BITS F423T

By

Sagnik Bhattacharya (2012B5A4329G)

under the supervision of Dr. Prasanta Kumar Das (Dept. of Physics) and

Dr. Sukanta Mondal (Dept. of Biological Sciences)

December, 2016

Acknowledgements

The efforts of my supervisors Dr. Prasanta Kumar Das and Dr. Sukanta Mondal are duly acknowledged. I would also like to thank Mr. Akshar Gupta and Ms. Shivangi Kamat for their help and support.

Certificate from the Supervisor

This is to certify that the Thesis entitled '***SIR model for infectious diseases***' is submitted by Mr. Sagnik Bhattacharya (ID: 2012B5A4329G) in partial fulfillment of the requirements of BITS F423T.

The thesis embodies the work done by him under my supervision.

Pramanta Kr. Das S. Srikanta Mandal

Signature of the supervisor

Dr. Pramanta kumar. Das. Physics dept · Associate Prof ·

Dr. Srikanta Mondal, Dept. of Biological Science. Assistant Prof.

Name and Designation

Date 8/12/16.

Abstract

Mathematical modeling is an important tool to study the spreading of infectious diseases. An important parameter in disease spread modeling is the ***basic reproduction ratio*** (R_0). Here, we have used a simple SIR (abbreviation of Susceptible-Infected-Recovered) model to estimate this parameter R_0 for 2 infectious diseases – Hong-Kong flu in New York, 1969 and EBOLA in Western Africa, 2014. The code written for the same is in Python, the *scikit-learn* library is used to obtain the best-fit curves from real data. Results reveal that – 1.) Hong-Kong flu is a non-epidemic and 2.) EBOLA is an epidemic.

Contents

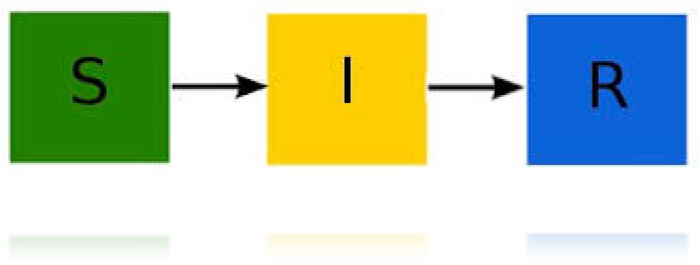
List of symbols

1. S, I, R : Susceptible, infected, and recovered population respectively.
2. β : Transmission coefficient.
3. γ : Recovery coefficient.
4. R_0 : Basic reproduction ratio.
5. N : Total population of the system.

<u>Introduction to the SIR disease Model</u>

The mathematical modeling of diseases began with Kermack and McKendrick studying the advent of the Bombay Plague in 1927. Since the 1970s, this field of interdisciplinary science has been put on a firm theoretical setting. The analysis of such a simple model, known as the SIR model (abbreviation for ***S***usceptible-***I***nfected-***R***ecovered) introduced by Kermack and McKendrick themselves. The SIR model is based on the following simplified assumptions:

1. The disease so concerned is contagious.
2. The disease affects a ***closed system*** i.e. there is no external influence on the system (a human population in this case).
3. The net population of the closed system is always conserved i.e. there are no births and deaths which occur in the time-frame wherein the disease is analyzed. This essentially means that the SIR model is useful only for the modeling of those diseases whose dynamics are much faster than the dynamics of births and deaths (like influenza, measles, mumps etc).
4. The population under consideration is classified into 3 types:
   a) **<u>Susceptibles:</u>** Those individuals of the population who haven't contracted the disease yet i.e. are yet to be infected.
   b) **<u>Infecteds:</u>** Those individuals who have contracted the disease and not yet recovered.
   c) **<u>Recovereds:</u>** Those individuals who have gone through the disease and have recovered from it. A recovered individual is assumed to become immune to the disease and is no longer a susceptible.
5. The flow involved in the model is :



[1]

i.e. a susceptible becomes an infected and then recovers. There is no other possible flow.

[1] Image from www.sherrytowers.com

The differential equations of SIR model

Let *S(t), I(t),* and *R(t)* denote the susceptibles, infecteds, and recovereds for a given population at any given time *t*.

Then, for any time *t*,

$$\frac{dS}{dt} = -\beta SI \qquad \boxed{(1)}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \qquad \boxed{(2)}$$

$$\frac{dR}{dt} = \gamma I \qquad \boxed{(3)}$$

Where *β* and *γ* are constants; *β* represents the rate at which a unit susceptible population converts to a unit infected population, *γ* similarly represents the rate at which a unit infected population recovers.

It can be noticed that for any time *t* and for any real *β* and *γ*, the following holds:

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0, \qquad \boxed{(4)}$$

Or in other words, as we discussed:

$$S(t) + I(t) + R(t) = \text{Constant} = N,$$

where, *N* is the total population that remains constant.

Let us now discuss the physical implications of the differential equations.

Equation (1) signifies that, for a population enduring a contagious disease, if there are *S* number of susceptibles and *I* number of infecteds, there are *S* x *I* possible routes of infection. Thus, the rate of infection ***is proportional to the number of possible routes of infection***.

Equation (3) is similar to the famous law of mass action of radioactivity. It says that more are the number of infected individuals, higher is the rate of recovery. Equation (2) thus can be derived from equations (1), (3), and (4).

Mathematically, the system of equations is non-linear and does not admit general analytical solutions. Nevertheless, significant results can be derived analytically.

An important parameter for analysis is the ***basic reproduction ratio*** R_0.

$$R_0 = \frac{\beta}{\gamma}.$$

Dividing equation (1) by equation (3), separating the variables and integrating further on, we get

$$S(t) = S(0)e^{-R_0(R(t)-R(0))}.$$

$$\boxed{(5)}$$

where S(0) is the initial number of susceptibles at time $t = 0$.

Also,

$$\frac{dI}{dt} = (R_0S - 1)\gamma I$$

$$\boxed{(6)}$$

From equation (6), we can deduce that if and only if:

$$R_0 S > 1$$

Then,

$$\frac{dI}{dt}(0) > 0$$

And there will be a proper disease outbreak. Thus, the initial condition deciding whether the disease will undergo an outbreak or not, is the ratio R_0.

For the disease to be classified as an epidemic, R_0 should be much greater than unity.

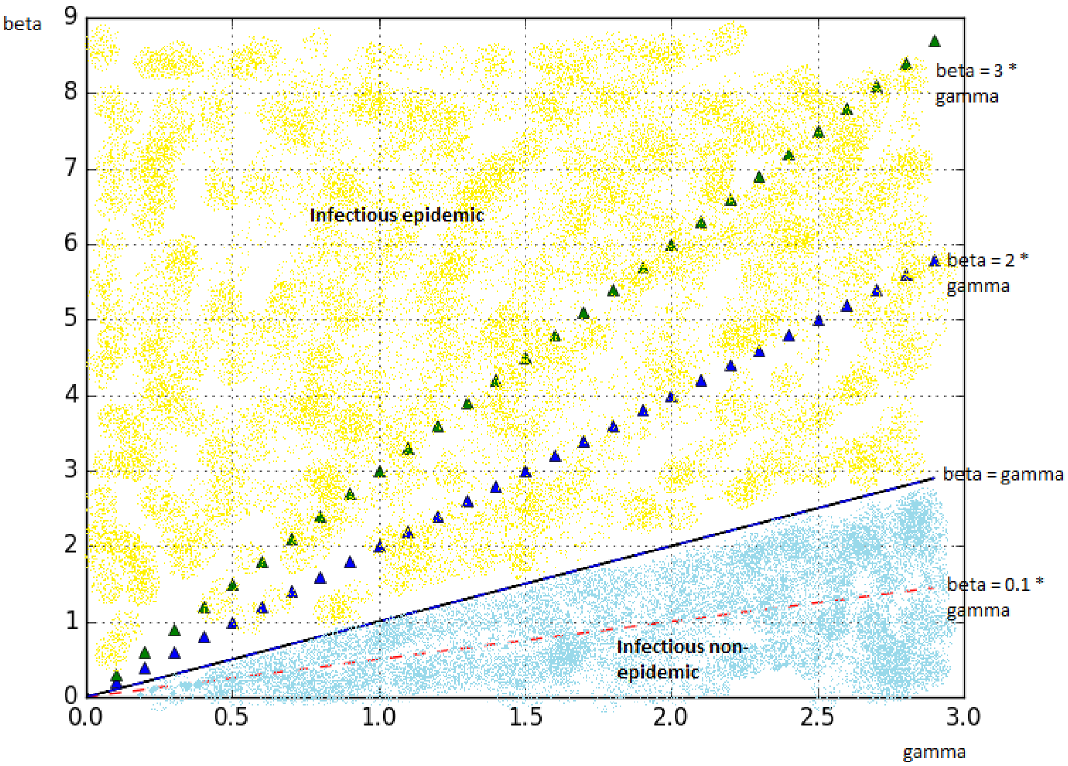


Fig. 1: Phase space diagram between β and γ showing the regions of infectious epidemic and non-epidemic diseases.

Analytical solutions to the SIR model

The susceptible, infected, and recovered populations can be analyzed using parametric solutions.

Differentiating equation (1) with respect to time yields:

$$\frac{dI}{dt} = \frac{-1}{\beta}\left[\frac{S''}{S} - \left(\frac{I'}{I}\right)^2\right]$$

$$\boxed{(7)}$$

On inserting (1) and (7) into (2), we get:

$$\frac{S''}{S} - \left(\frac{S'}{S}\right)^2 + \gamma\frac{S'}{S} - \beta S' = 0$$

$$\boxed{(8)}$$

Also, eliminating *I* from equations (1) and (3) gives:

$$\frac{dR}{dt} = -\frac{\gamma}{\beta}\left(\frac{S'}{S}\right)$$

$$\boxed{(9)}$$

which can be integrated to give:

$$S = C_0 e^{\frac{-\beta}{\gamma}R}$$

$$\boxed{(10)}$$

where C_0 is any positive integration constant.

The value of S at $t = 0$ provides the following value for the integration constant:

$$C_0 = S(0).\, e^{\frac{-\beta}{\gamma}R(0)}$$

$$\boxed{(11)}$$

Now, from equation (10), we have :

$$dS/_{dt} = -C_0\, {}^{\beta}/_{\gamma}\, R'\, e^{\frac{-\beta}{\gamma}R}$$

$$\boxed{(12)}$$

Now, differentiating equation (9), we obtain the second order Ordinary Differential Equation:

$$\frac{d^2R}{dt^2} = -\frac{\gamma}{\beta}\left[\frac{S''}{S} - \left(\frac{S'}{S}\right)^2\right]$$

$$\boxed{(13)}$$

Now, inserting equations (9), (12), and (13) to (8) , we obtain the basic differential equation which describes the dynamics of spread of a non-fatal disease in a given closed population.

$$d^2R/_{dt^2} = C_0\beta R' e^{-\frac{\beta}{\gamma}R} - \gamma R'$$

(14)

Now, we start the process of parameterization in order to get a better picture of what the above differential equations represent.

Let us define a new function *W*(*R*) parameterized by the variable *R*:

$$W = e^{-\frac{\beta}{\gamma}R}$$

(15)

At *t=0* , *W* has the initial value

$$W(0) = e^{-\frac{\beta}{\gamma}R(0)}$$

(16)

Now, substituting (15) to (14) yields the second order Ordinary Differential Equation for W:

$$W\,{}^{d^2W}\!/_{dt^2} - \left({}^{dW}\!/_{dt}\right)^2 + (\gamma - \beta C_0 W)W\,{}^{dW}\!/_{dt} = 0$$

$\boxed{(17)}$

We further define another variable ρ:

$$\rho = {}^{dt}\!/_{dW}$$

$\boxed{(18)}$

With the help of this transformation, (16) can be expressed as a Bernoulli type differential equation:

$$\frac{d\rho}{dW} + \frac{\rho}{W} = (\gamma - C_0\beta W)\rho^2$$

$\boxed{(19)}$

The subsequent general solution is thus given by:

$$\rho(W) = {}^{1}\!/_{W(C_1 - \gamma \ln W + C_0 \beta W)}$$

$\boxed{(20)}$

where C_1 is an arbitrary constant of integration.

Simultaneously, we can obtain the integral representation of time as:

$$t - t_0 = \int_{W_0}^{W} dw \Big/ w(C_1 - \gamma \ln w + C_0 \beta w) \qquad (21)$$

where t_0 is another arbitrary constant of integration.

We have now thus arrived with the complete set of equations in the parametric form.

$$S = C_0 W \qquad (22)$$

$$I = \frac{\gamma}{\beta} \ln W - C_0 W - \frac{C_1}{\beta} \qquad (23)$$

$$R = -\frac{\gamma}{\beta} \ln W \qquad (24)$$

Adding (22), (23), and (24) should yield the total population *N*.

Thus, we have

$$C_1 = -\beta N$$

(25)

Advantages and disadvantages of the SIR model

Advantages:

1. The model is quick and easy to understand, even though exact analytical solutions don't exist.
2. Ever since the model was proposed, it has been widely used by the scientific community to explain the dynamics of disease/epidemic spread.
3. Lesser number of variables in the system of equations makes the solving of SIR model computationally easy.
4. Significant accuracy can be achieved in its simulation when the time interval is taken to be very small.

Disadvantages:

1. It is difficult to model epidemic scenarios with a very high death rate using SIR model. It is only accurate for small isolated populations with uniform population density.
2. Estimation of transmission parameters like β and γ is a difficult task. The model is highly sensitive to these parameters and a small change in these leads to deviation of the trajectories by a large extent.

Simulation of SIR model subjected to boundary conditions

Now, we begin our simulation of the mathematical model using Euler's method (a Runge-Kutta numerical method of first order).

For the following various initial conditions, we have derived the corresponding various plots to check how sensitive they are to changes in β and γ.

1. $\beta = 0.06$, $\gamma = 0.008$; $S(0) = 56$, $I(0) = 6$, $R(0) = 0$; $R_0 = 7.5$

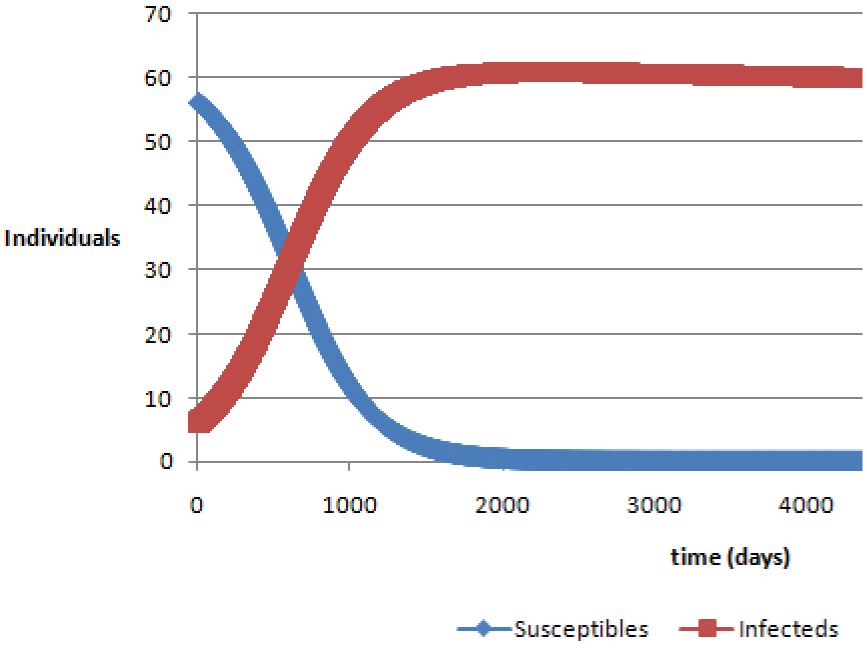


Fig. 2: Susceptibles v/s Infecteds simulated for a duration of 4000 days.

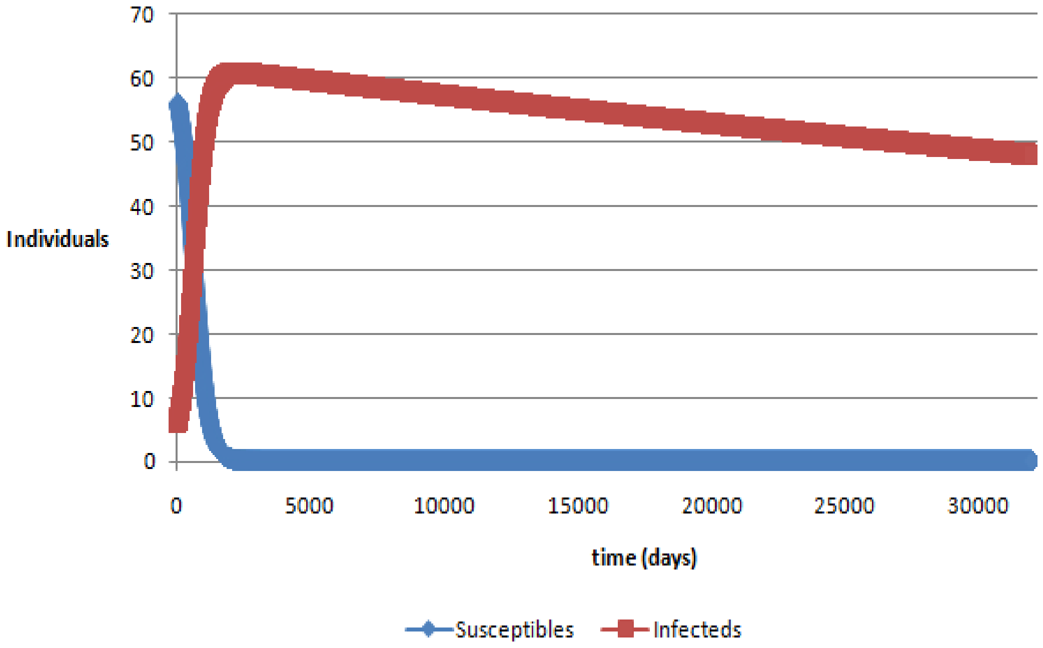


Fig. 3: Susceptibles v/s Infecteds simulated for a duration of 30000 days.

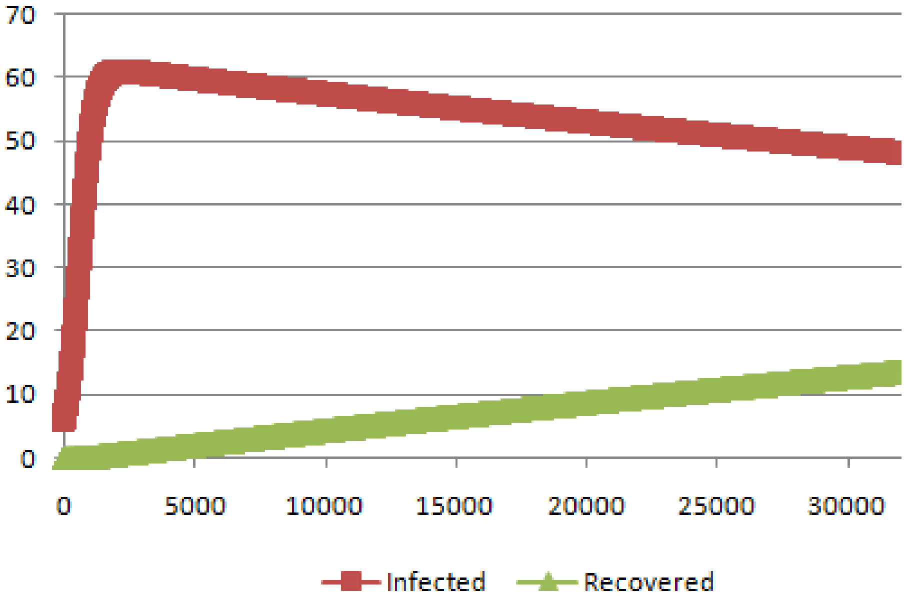


Fig. 4: Infecteds v/s Recovered simulated for a duration of 4000 days.

2. $\beta = 0.065$, $\gamma = 0.008$; $S(0) = 56$, $I(0) = 6$, $R(0) = 0$; $R_0 = 8.125$

Here, we have reduced the value of β in order to check the sensitivity of the plot with respect to β only.

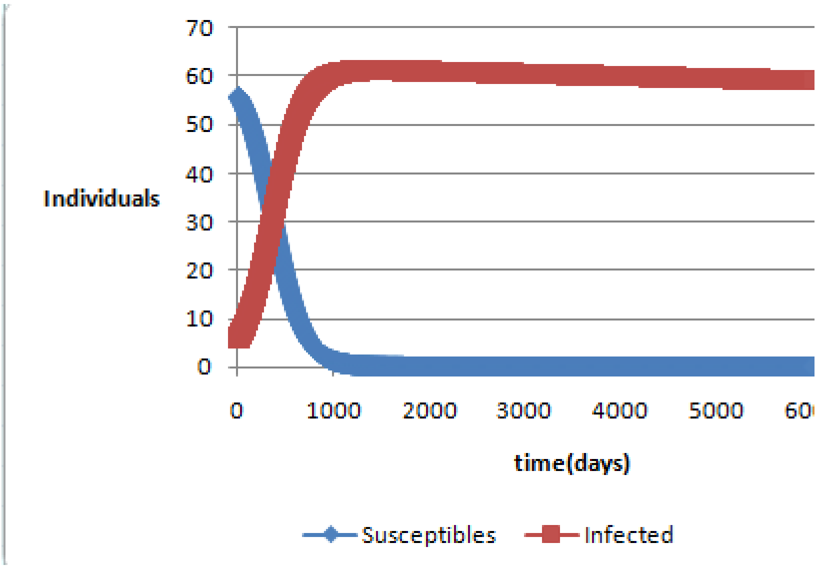


Fig. 5: S and I versus time for $\beta = 0.065$

Comparing this with the previous plot ($\beta = 0.06$)

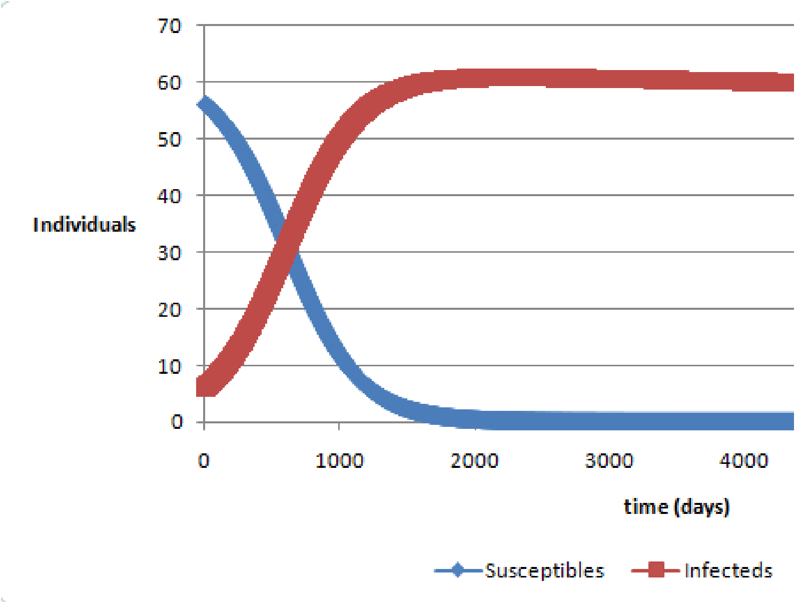


Fig. 6: S and I versus time for $\beta = 0.060$

From figure 5, we can see that increasing the value of β drastically increases the steepness of the rate of infection and the rate of drop of susceptibility.

3. $\beta = 0.065$, $\gamma = 0.001$; $S(0) = 60$, $I(0) = 7$, $R(0) = 0$; $R_0 = 65$
   $\beta = 0.065$, $\gamma = 0.0025$; $S(0) = 60$, $I(0) = 7$, $R(0) = 0$; $R_0 = 26$
   $\beta = 0.065$, $\gamma = 0.008$; $S(0) = 60$, $I(0) = 7$, $R(0) = 0$; $R_0 = 8.125$
   are compared here to check sensitivity on γ. (figures 7 to 9)

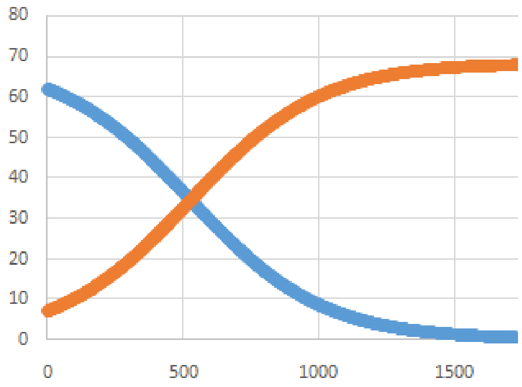


Fig. 7: S and I versus time for $\gamma = 0.001$

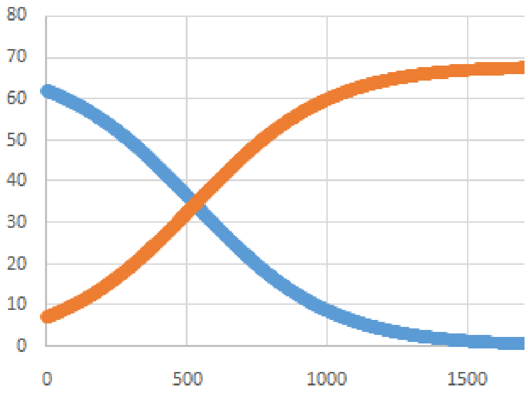


Fig. 8: S and I versus time for $\gamma = 0.0025$

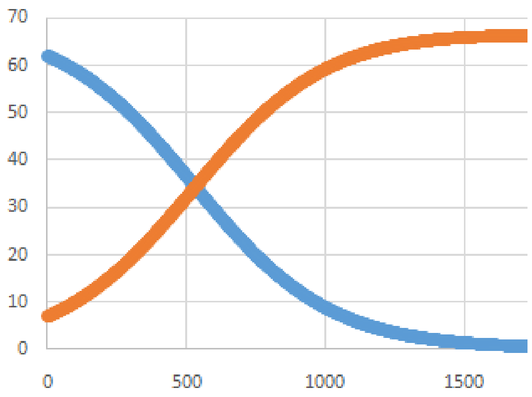


Fig. 9: S and I versus time for $\gamma = 0.008$

Here, we can see that changing γ significantly does not produce a change in the above analyzed region of curve of infecteds and susceptibles. The effect is only pronounced in the recovery v/s infecteds curve after a long time, as I and S decrease further and further, so does the product S x I . The effect of β is on the other hand far more impactful in the above analyzed region, due to the product of S and I being high.

The following are plots (figures 10 to 12) of recovered populations corresponding to different γ values. Here we see that γ has a significant role due to the direct relation with the number of recovered individuals R.

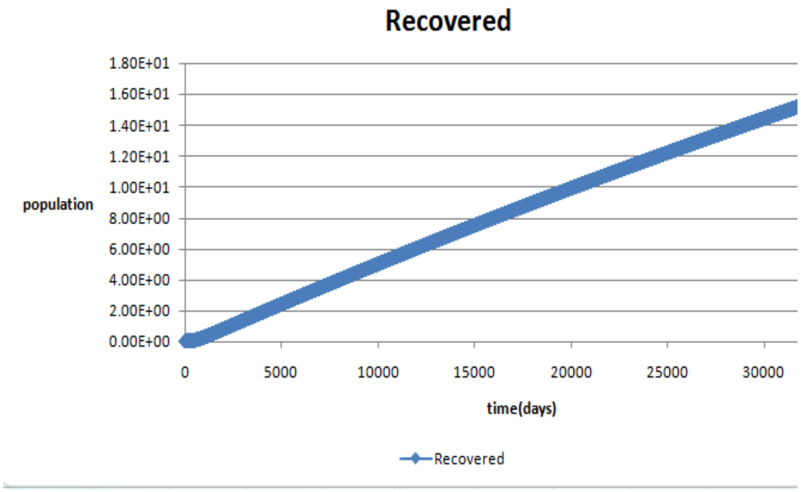


Fig. 10 : R versus time for $\gamma = 0.001$

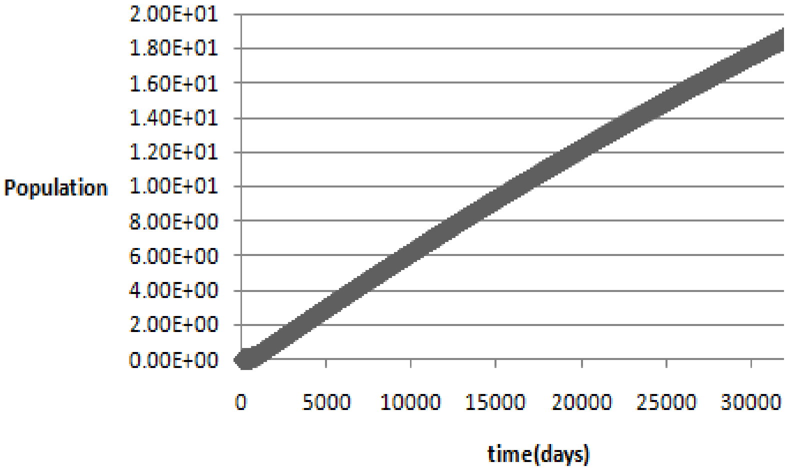


Fig. 11: R versus time for $\gamma = 0.0025$

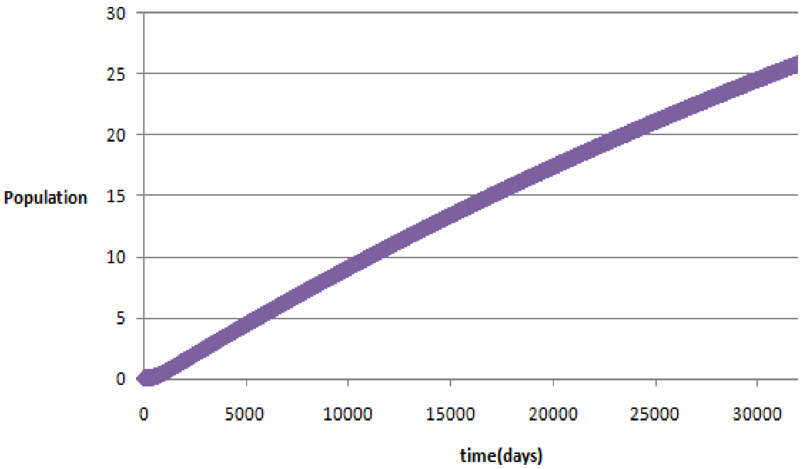


Fig. 12 : R versus time for $\gamma = 0.008$

Testing SIR model to a disease: Hong-Kong flu in New York, 1968-69

During the winter of 1968-69, a virulent disease named *Hong-Kong* flu swept through the United States. Since it was a newly discovered disease, vaccination was not available at that time.

We will use a data sample which was obtained in this context, and will try to estimate the fitting of the data to the SIR model. Here it is assumed that the number of infected individuals corresponds to the number of cases reported per week and the time below is given in weeks. The table below (Table 1) gives the number of infected individuals measured v/s time in months (refer 3rd reference in bibliography).

Table 1: Cases of Hong-Kong flu in Manhattan, New York (1969).

Time	***Infected individuals***
1	1
2	6
3	16
4	54
5	150
6	442
7	582
8	510
9	475
10	383
11	251
12	140
13	56
14	24
15	10
16	0

Polynomial interpolation assumes that the dependent variable is of a polynomial form with respect to the independent variables and subsequently we are required to find the coefficients of this polynomial by tallying with available data.

A polynomial interpolation (up to 6th order polynomial), using least squares minimization without regularization gives the following fit:

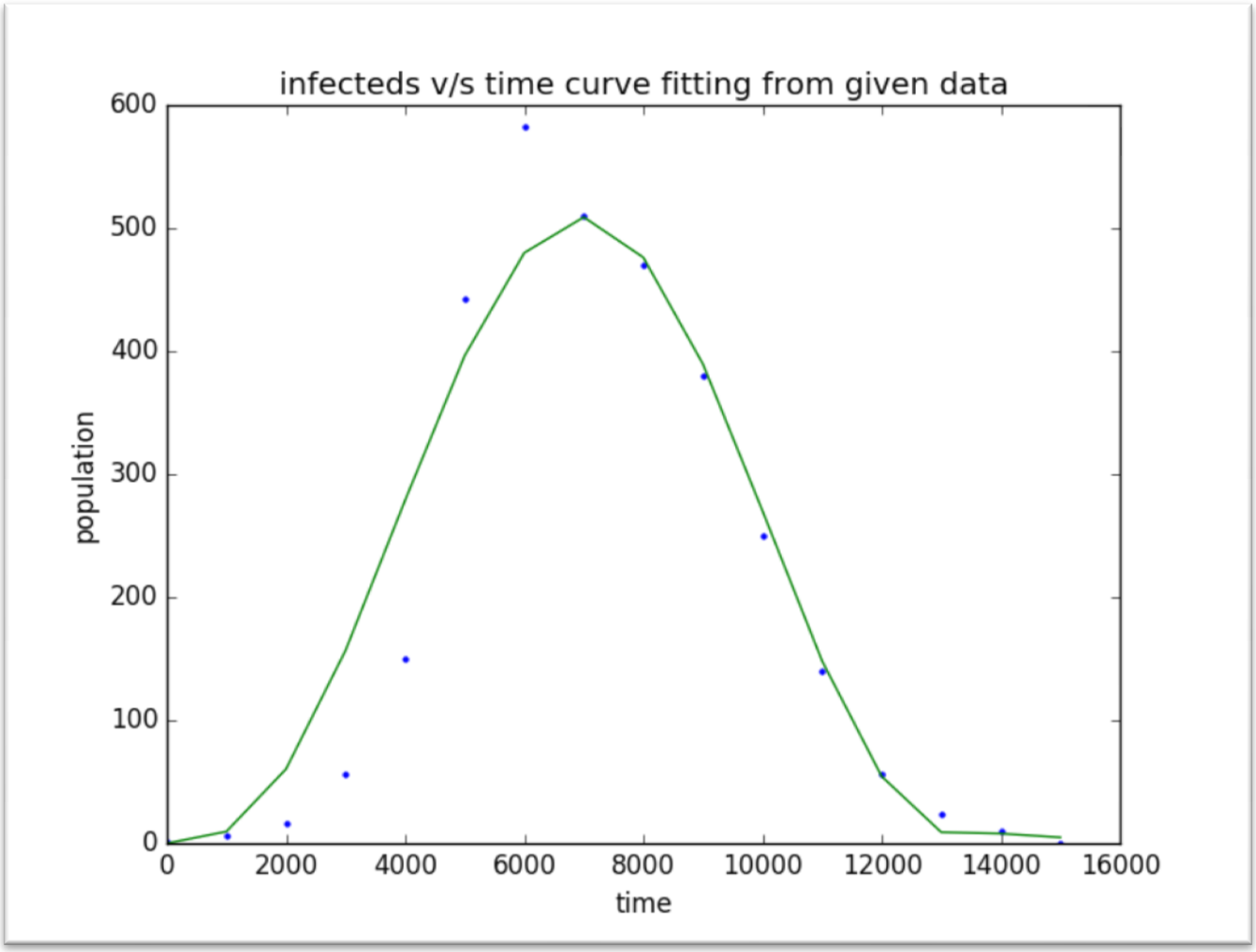


Fig. 13: Polynomial interpolation upto 6th order polynomial using least squares minimization without regularization. Time is represented in units of one thousand. The dots represent the actual data points given in the table.

The procedure of least-square fitting gives an optimum fit for a specific order of polynomial expansion up to 6th order polynomial.

The optimum coefficients of the polynomial are subsequently computed:

```
-4.395250E-21   -1.844396E-16   -2.578250E-12

1.199968E-4      5.2335054E-12   1.540177E-15

4.027991E-19
```

The reproduction ratio R_0 is thus evaluated to be 0.67 which is less than unity, and thus Hong-Kong flu does not classify as an epidemic. It should be noted that the value of R_0 is measured at time $t = 0$ and not the time-averaged value. This is because over the course of time, the value of R_0 changes as per measurements from the data points. This gives a spectrum of possibilities for the values of R_0. However, small changes in R_0 lead to significant changes in the plot of S, I, and R versus time. Hence, it is imperative to understand that the value of R_0 has its physical relevance during the initial phase of the outbreak.

In this case, a 6th order polynomial expansion gives the most ‘intuitive’ fit, a further increase (i.e. 7th order) leads to too much sensitivity on each and every data point and hence the curve undulates more compared to the previous curve (Fig. 13).

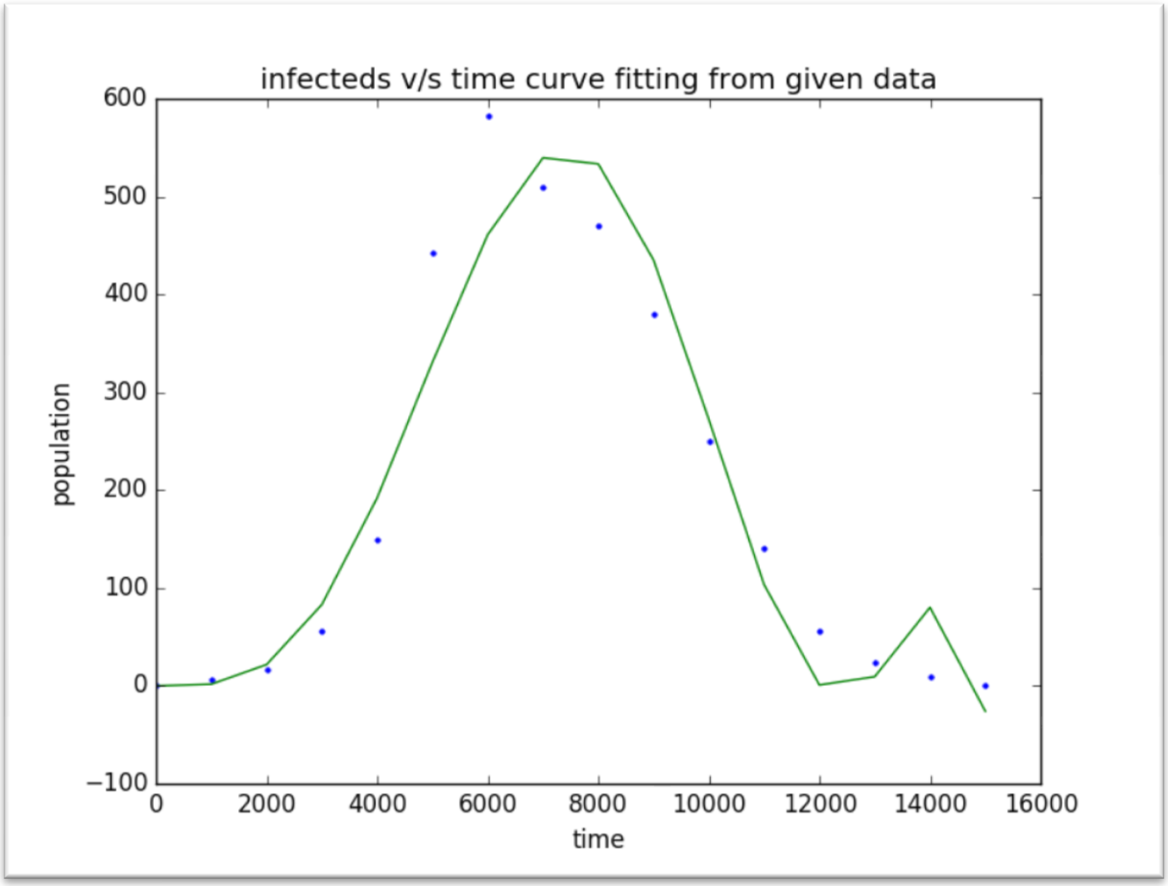


Fig 14: Polynomial fit for 7th order polynomial interpolation.

Similarly, a 5th order fit is not sufficient enough in terms of accuracy.

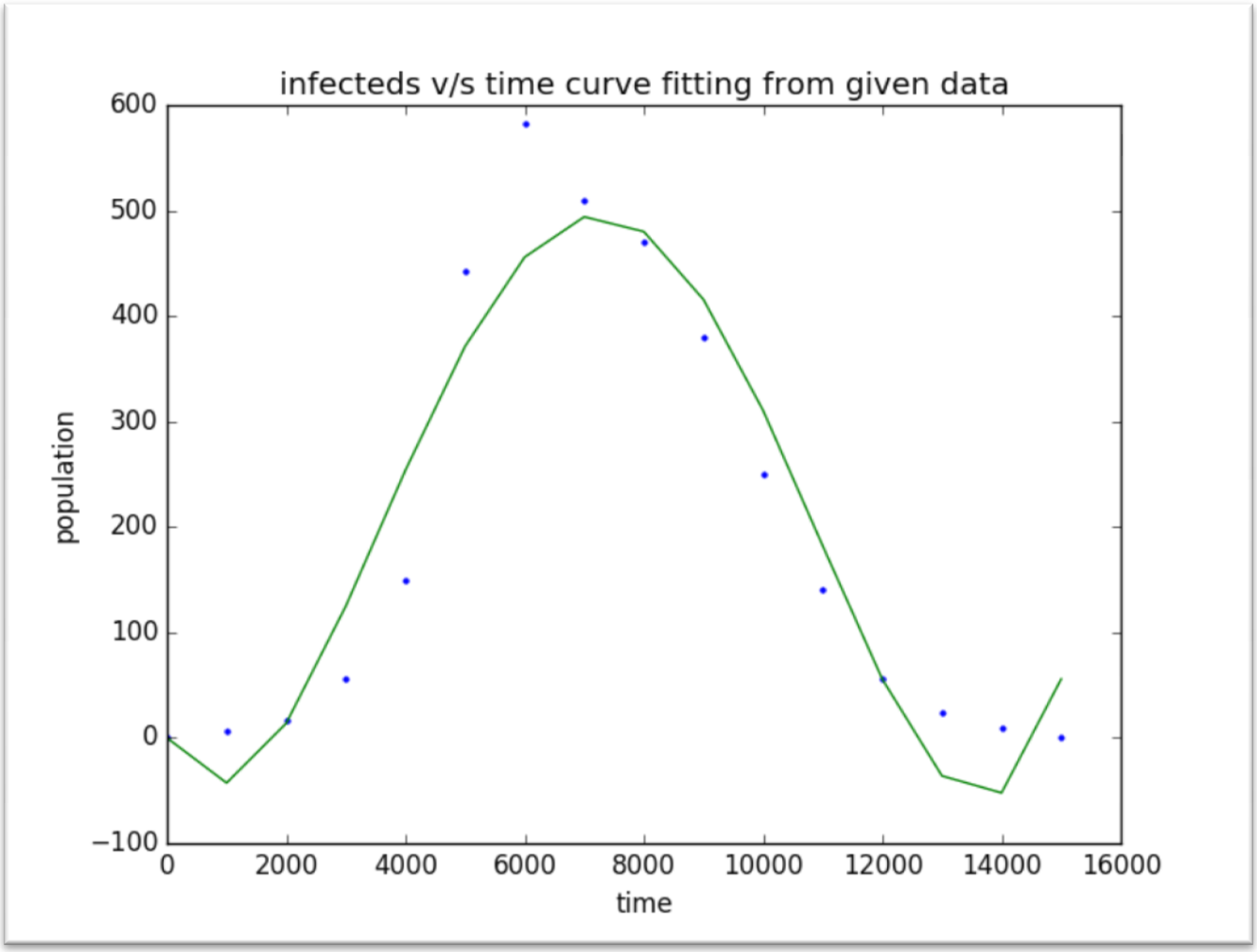


Fig. 15: Polynomial fit for 5th order interpolation.

Due to the inherent trade-off between fitting accuracy and trend/ prediction (known as the bias-variance trade-off), the phenomenon of over-fitting occurs, and thus the polynomial interpolation model without regularization is not deemed fit for prediction purposes.

In order to deduce trend from data for future prediction purposes, it is better to use fitting algorithms with regularization.

With ridge-regularization (also known as Tikonov regularization), the same set of data yields another fit (value of regularizer coefficient λ = 1.0). It can be observed that the curve obtained is smoother, this time for a 2nd degree polynomial interpolation. However, there is a visible compromise on accuracy, especially at the end-points, which is as expected.

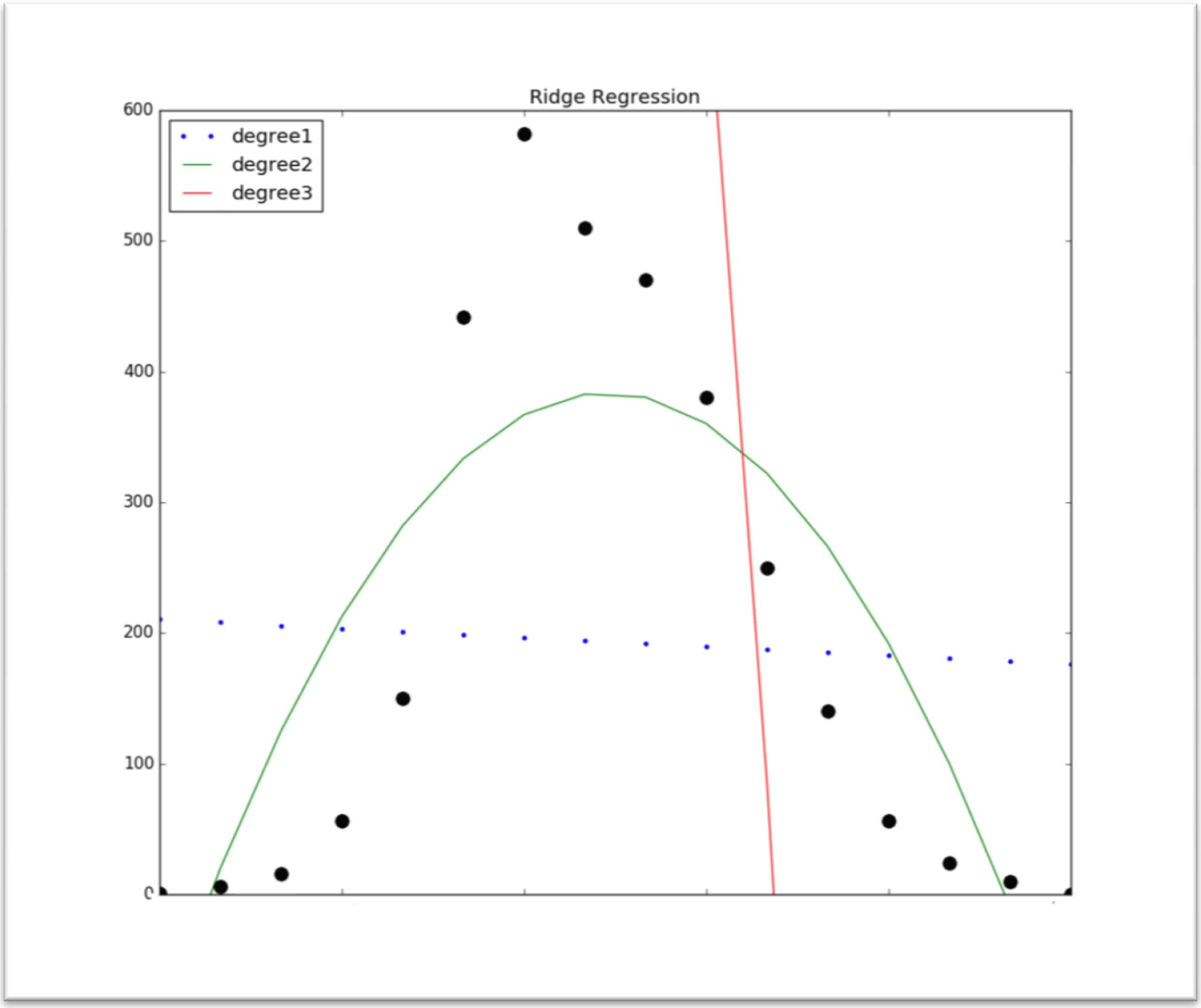


Fig. 16: Polynomial interpolation with ridge regularization for the same data (λ = 1.0). The blue dotted line represents polynomial interpolation up to first order, the green line represents the same up to 2nd order, and the red line represents the same up to 3rd order.

The coefficients for the 2nd order fit are computed:

```
-0.002018              1.983354          1.453666E-22
```

Testing SIR model to a disease: EBOLA in Sierra Leone, 2014

Similarly, we now analyze the SIR model to data corresponding to the outbreak of EBOLA epidemic in West Africa. The number of reported cases in Sierra Leone v/s time in days is plotted along with the best fit curve (figure 16). (Refer 4th reference in bibliography for more details on the data.)

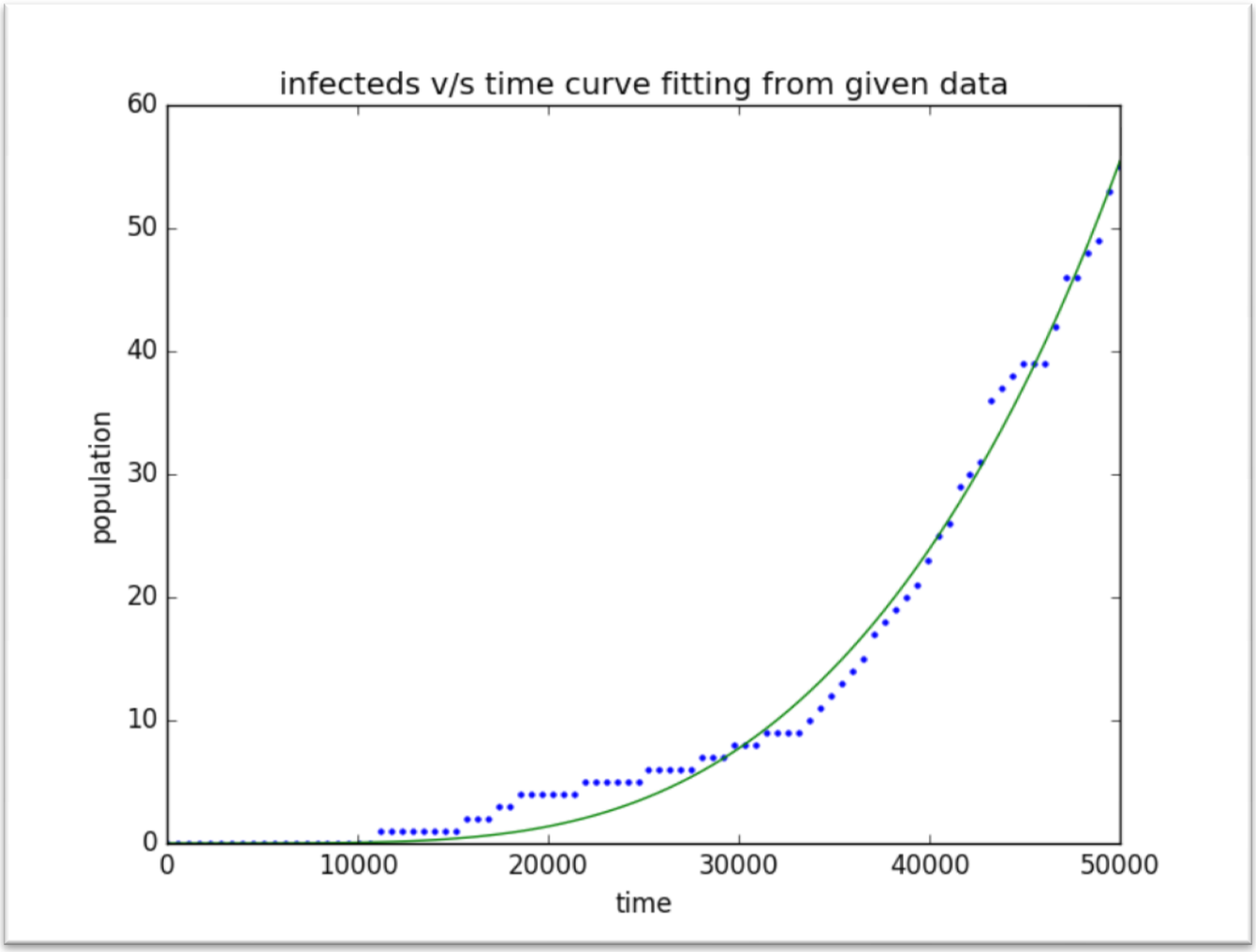


Fig. 17: Number of cases v/s days during the EBOLA outbreak at Sierra Leone, 2014. Days should be scaled down 1000 times.

The above plot (figure 17) is obtained through 7th order polynomial interpolation without regularization.

The coefficients obtained for the same are as follows:

```
1.392457E-31    -1.813513E-26       7.360508E-22

5.711422E-26    3.030242E-30        1.381619E-34

5.890292E-39    2.460973E-43
```

The basic reproduction number R_0 thus calculated is of the value 2.65. Hence, the EBOLA disease spread in Western Africa in 2014 can be classified as an epidemic. This is close to the reported value of 2.53 (see reference 9 in bibliography).

Conclusion

The SIR single strain model was analyzed and its analysis was extended to 2 cases of diseases i.e. Hong-Kong flu in New York, 1968-69 and EBOLA in Sierra Leone, 2014.

Further scope of analysis exists in simulating epidemic through network models like Neural Networks. Individual-specific parameters like average time duration for which an individual is infected can be obtained through these methods.

Bibliography

1. "Exact analytical solutions of the SIR epidemic model with equal death and birth rates" by T. Harko, F.S.N. Lobo, and M.K. Mak.
2. "Non-linear dynamics and chaos" by S. Strogatz.
3. "Seasonal Influenza in the United States, France, and Australia: Transmission and prospects for control" by G. Chowell, M. Miller, and C.Viboud.
4. "Modeling EBOLA in West Africa: Cumulative cases by date reporting" by *Contagious Disease Surveillance*, 2014.
5. "SIR model for spread of disease- the differential equation model" by D. Smith and L. Moore, *Mathematical Association of America*.
6. "SIR model", *en.wikipedia.org/wiki/Epidemic_model*.
7. "Pattern Recognition and Machine Learning" by C. Bishop.
8. "Introduction to Machine Learning in Python with scikit-learn", *ipython-books.github.io/featured-04*
9. "Estimating the reproduction number of EBOLA virus during the 2014 outbreak in West Africa", C. Althaus.

Appendix

The following Python code is used for generating the SIR plots as well as curve-fitting from real data for the case studies mentioned before.

```
import numpy as np
import math
import matplotlib.pyplot as plt
import pylab as p

##initial conditions
h = 0.001
duration = 1000
N = int(duration/h)
t = np.linspace( 0.0 , duration , N )

R0 = 192
beta = 0.0026
gamma = beta/R0
susceptible = np.array( [ 0.0 for x in xrange( N ) ] )
infected = np.array( [ 0.0 for x in xrange( N ) ] )
recovered = np.array( [ 0.0 for x in xrange( N ) ] )
susceptible[ 0 ] = 562.0
infected[ 0 ] = 1.0
recovered[ 0 ] = 1.0
N = susceptible[ 0 ] + infected[ 0 ] +  recovered[ 0 ]

##Define the system of ODEs
def dS_dt( S , I  ) :

	return  - 1 * beta * S * I

def dI_dt( S , I  ) :

	return  ( beta * S * I ) - ( gamma * I )

def dR_dt( I ) :

	return  gamma * I

def solve_SIR():

	# Solve ODES via Numerical Integration

	for i in range(1 , N) :

		susceptible[ i ] = susceptible[ i - 1 ] + ( dS_dt( susceptible[ i - 1 ] ,
infected[ i - 1 ] ) * h )
		infected[ i ] = infected[ i - 1 ] + ( dI_dt( susceptible[ i - 1 ] ,
infected[ i - 1 ] , recovered[ i - 1 ] ) * h )
		recovered[ i ] = recovered[ i - 1 ] + ( dR_dt( recovered[ i - 1 ] ) * h )

	P = np.array([ susceptible , infected , recovered ])

	Return P
```

```
def fit_SIR():

        solve_SIR()

        #polynomial interpolation without regularization
        X_data = np.linspace( 0 , 15000 , 16 )
        Y_data = np.array( [ 1, 6 , 16 , 56 , 150 , 442 , 582 , 510 , 470 , 380 , 250 ,
140 , 56 , 24 , 10 , 0 ] )

        degree = 5

        A = np.vander( X_data , degree )
        ( coeffs , residuals , rank ,  sing_vals ) = np.linalg.lstsq( A , Y_data )

        f2 = np.poly1d( coeffs )
        Y_estimate = f2( X_data )

        plt.plot( X_data , Y_data , '.' , label =      'original data' , markersize = 5
)
        plt.plot( X_data , Y_estimate , '-' , label =  'estimate data' , markersize = 3
)

        plt.xlabel( 'time' )
        plt.ylabel( 'population' )
        plt.title( 'infecteds v/s time curve fitting from given data' )
        plt.savefig( 'SIR_fit.png' )

        print "degree of vandermonde matrix is" , degree
        print coeffs




##the following curve fittings are for Hong-Kong flu, New York 1968-69.


def ridge():

        solve_SIR()

        #ridge regularization of data
        X_tr = t



        x = np.linspace( 0.0 , 1000 , 16 )
        y = np.array( [ 1 , 6 , 16 , 56 , 150 , 442 , 582 , 510 , 470 , 380 , 250 , 140
, 56 , 24 , 10 , 0 ] )

        plt.figure( figsize = ( 14 , 14 ) )
        plt.xlim( 0 , 1000 )
        plt.ylim( 0 , 800 )

        ridge = lm.RidgeCV()
        ridge.alpha = 1.0

        for deg , s in zip( [ 1 , 2 ,  3 ] , [ '.' , '-' , ‘r-‘ ] ) :

                ridge.fit( np.vander( x , deg + 1 ) , y )
                y_ridge = ridge.predict( np.vander( X_tr , deg + 1 ) )
                plt.plot( X_tr , y_ridge , s , label = 'Degree' + str( deg ) )
        plt.plot( x , y , 'ok' , ms = 10 )
        plt.title( 'Ridge Regression' )
        plt.savefig( 'SIR2.png' )
        print ridge.coef_

##End of code.
```