Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Ann Njambi Karong'oi

COMPARING DIFFERENT APPROACHES OF EPIDEMIOLOGICAL MODELING

Master's Thesis

Examiners: Professor Heikki Haario
Professor Matti Heiliö

Supervisor: Professor Heikki Haario

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Ann Njambi Karong'oi

Comparing Different Approaches of Epidemiological Modeling

Master's Thesis

2018

44 pages, 13 figures and 3 table.

Examiners: Professor Heikki Haario
Professor Matti Heiliö

Keywords: deterministic, discrete, modeling, distributions, models, compartments

Disease prevention, control and eradication has remained a major challenge for centuries. Epidemiological modeling is a key and powerful tool that plays a role in establishing integrated control methods that can be applied to lessen the disease burden and reduce mortality rates especially during epidemics. Different kind of epidemiological models are developed depending on the dynamics of the disease and are mostly compartmental models.

This thesis aims to compare different modeling approaches that can be applied to epidemiological models to yield results that are close to reality. In particular, we concentrate on the discrete modeling rather than the classical ODE based modeling that has some drawbacks that makes it inefficient in some cases of disease modeling. We also double check if the waiting times in the model compartments can be said to come from any other distribution rather than the exponential distribution.

PREFACE

My most sincere gratitude to the almighty God for His blessings, grace and the unending love.

Special thanks and appreciation to my supervisor Professor Heikki Haario for the huge support, guidance and insightful comments throughout this work. To my family, words can't be enough to thank you for the support, prayers and the encouragement. To my great friends Eddymurphy and Ann Chepkoech, you are an inspiration to me, thank you for the motivation and kind words of encouragement at all times.

Finally, I thank the entire Lappeenranta University of Technology family, administration, lecturers, tutors, staff, and my fellow students, you made life at LUT interesting and worth living, I appreciate all of you. May Almighty God bless you all abundantly.

Lappeenranta, May 25 2018

Ann Njambi Karong'oi

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS AND SYMBOLS

ODE	Ordinary Differential Equations
LSQ	Least Squares
SEIR	Susceptible-Exposed-Infected-recovered
SARS	Severe Acute Respiratory Syndrome
θ	Parameter estimates.
ϵ	measurement error.
$\boldsymbol{X}$	a matrix
R_0	Basic Reproduction Number

1 INTRODUCTION

1.1 Background

Epidemiology is the study of the distribution of diseases, their causes and the prevalence in a population of susceptible.

Since the first use of mathematical concepts to explain the spread of smallpox by Daniel Bernoulli in 1970, the theory of epidemiology has become popular and widely used. The idea of deterministic epidemiological modelling is dated back in the 20^{th} century. Hamer in 1906 developed a discrete model with the aim of studying the re-occurrence of measles. In his model, he made an assumption that the incidence of the disease depends on the product of the number of susceptible and infected humans. Ross in 1911 developed a deterministic model to study the dynamics and the control measures for malaria. He considered malaria to be a host-vector disease [1].

Epidemiological modeling has been a major tool in influencing critical decisions pertaining the development of health policy, implementation of control measures for the infectious diseases as well as the prevention. It aids the public health decision makers with knowledge of how a multidisciplinary approach can be implemented with the aim of reducing the disease burden in a population.

Today, epidemiological modelling has been playing a key role in providing information on the most efficient and cost effective strategy of fighting infectious diseases. It has been instrumental in predicting if the implementation of the possible solutions can reduce the number of infective [2].

A model is a mathematical representation of a system that is designed to increase and enhance the understanding of the system in question. Epidemiological models represent the underlying relationship of the system components, how they interact and gives an idea of how the system would behave overtime. These helps in understanding the effect of the external factors on the output.

Different types of mathematical models are used in modelling. There are simple ones which are expressed in terms of differential equations commonly referred to as deterministic and more complex discrete models which are stochastic in nature. Spatial models put more emphasis on distances and locations when used to study the disease spread and transmission. They concentrate more on the role of geographical factors in disease spread and evaluate control measures which are spatially targeted [3].

Various categories of models depend on different aspects being considered. In terms of time, the model is either discrete or continuous. Another aspect is the structure of the population involved, whether it is homogeneous or heterogeneous. In a homogeneous population structure, members are assumed to have equal chances of being infected. On the contrary, a heterogeneous mixing assumes different classes within a population have unequal risks of contacting the infection [3].

The choice of the approach used depends on various factors; availability of data and its reliability, the knowledge of the dynamics of the disease being modelled, the modellers experience in the field of disease modeling.

Overtime, the emerging of new technology and computers has positively impacted the people's health. Recently, more sophisticated disease models are being developed and computers simulations in epidemiology are being done. This has made it easier to come up with more advanced solutions to complicated problems hence leading to improvement of the health level [2].

There are different types of compartmental models with different number of compartments that represent different state variables, often referred to as classes. Such models include ; SI, SIR, SEIR, SEIRS. The simple (SI) epidemic models consist of two classes of humans, the susceptible and infected. These type of models do not include the latency period. Once susceptible humans become infected, they remain in that infected forever. SIR models comprises of susceptible, infected and recovered humans respectively. In this case, when a susceptible human becomes infected, he or she can recover or die from infection. The SEIR models are similar to the SIR except that they consist of an exposed class which is the latency stage.

In the SEIRS compartmental models, the immunity acquired by the population in class R is considered to be temporary. This implies that the individuals are susceptible to the infection at the end of the temporary immune [1].

The choice of the compartments to be included in a model is influenced by the dynamics of the disease being modelled and what problem the model is expected to solve.

1.2 Objectives of the study

The aim of this thesis is to compare different approaches of modeling so as to see the one that gives us more realistic results as well as to see how other different methods are as compared to the traditional deterministic method. In addition, the basic ODE modeling only compares exponential waiting times, we therefore aim to use different distributions to study the effect of waiting times in modeling. Waiting times are the average times an individual spends in a given compartment before moving to another compartment. For an SEIR that we shall use, we have two types of waiting times to consider. Average time that a human stays in the exposed class(latent period) and the verage time that a human stays in the infected class before recovery.

Therefore, we concentrate on achieving the following outlined tasks:

- To understand the deterministic approach in modeling as well as its drawbacks.
- Implementation of discrete modeling using real data from a case study.
- To compare the results from the usual deterministic modeling to those of discrete approach in order to see the connection.
- To use different values as our initial infected population in order to observe the amount of variability caused especially in the case of the real death example.
- To describe the Gillespie algorithm as a discrete approach in modelling.
- To examine the effects of different latent time distributions.

1.3 Structure of the Thesis

In this section we give a brief overview of how the the tasks done in this study are arranged. This thesis contains six chapters. Chapter 1 comprises of the background of the study which gives a brief introduction of epidemiological modeling and and it's importance in the prevention and control of diseases. Here we also have the objectives of the study outlined. Chapter 2 consists of deterministic compartmental models. In particular we concentrate on the SEIR model which has been used to model the Influenza data. Deterministic modeling approach has been explained and also it's limitations have been discussed. A case study of San Francisco data has been used to model the infected using the SEIR model. Discrete modeling approach has been explained in Chapter 3. Here we talk of the discrete models and what makes them more suitable in epidemiological modeling. Two methods of discrete modeling has also been discussed are Gillespie algorithm and Euler method. In Chapter 4, we have introduced and discussed some probability distributions that have been used to model the waiting times in this thesis. The results from numerical simulations using four different distributions have been given and described in Chapter 5 and a conclusion of the work given in Chapter 6.

2 Deterministic Compartmental Models

2.1 SEIR Compartmental Models

SEIR models are based on a case where there are susceptible individuals who are at a risk of contacting an infection from an infectious agent. They revolve around the dynamics of the disease and changes occurring overtime when a small group of individuals become infected and start transmitting the infection(infectious).

The population is divided into four epidemiological classes known as the state variables. S represents those who are susceptible to the infection, E denotes individuals who are exposed to the infection but can not transmit it. I is the compartment of the infected humans who are infectious and can spread the infection. R comprises of those who have been removed either by recovering and gaining immunity such that they can not be re-infected, disease induced death or through isolation. Individuals in R can not contribute to disease spread [4].

The key question in epidemiological modelling is whether or not an infectious disease can invade a population of susceptibles causing an epidemic. This is often explained interms of a threshold value commonly referred to as the basic reproduction number R_0 [5].

The basic reproduction number R_0, is defined as the average number of secondary infections produced by a single infected individual during his/her period of infectiousness in an entire population of susceptibles [6]. The interpretation of the implication of this basic reproduction number can be explained as follows: for $R_0 < 1$, a single infected individual on average generates less than one new case of an infected individual prompting that soon the infection will die out in a population. If $R_0 > 1$,the infection will invade a population of susceptible individuals.

The density of the total population is assumed to be a constant and the population at time t is denoted by N(t) which is expressed as shown by (1) below;

$$N(t) = S(t) + E(t) + I(t) + R(t) \tag{1}$$

SEIR model is a compartmental model initial value problem given by the ODE system

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\alpha\, S\, I}{N} \\
\frac{dE}{dt} &= \frac{\alpha\, S\, I}{N} - \beta\, E \\
\frac{dI}{dt} &= \beta\, E - \gamma\, I \\
\frac{dR}{dt} &= \gamma\, I
\end{aligned}$$

where the $S(t), E(t), I(t)$ and $R(t)$ represents the susceptible, exposed, infected and removed population at any time t respectively.

For any initial conditions of our system we require that $S(t) > 0, E(t) \geq 0, I(t) \geq 0, R(t) \geq 0$ for all t.

2.2 Deterministic Modeling and its Limitations

Deterministic epidemic models are based on the idea of characterizing an epidemic by the susceptible individuals and the infected in a population over time. They are suitable when when the populations being modeled are large and where the model parameters are sufficient for an average individual [7].

The term deterministic can be defined as the expected trend when there is no randomness. The output from a deterministic model depends on the input such as the initial conditions and data and we always obtain same result for a particular set of input. In deterministic models, there is a continuous change in the state variables.

Although this models are often used to model infectious diseases, there are some limitations associated with them. One is the law of mass action assumption. Deterministic models assume a homogeneous mixing in population where by all individuals are assumed to have equal chances in the risk of becoming infected. The disadvantage of the

mass action assumption was evident in modeling the pandemic of Severe acute Respiratory Syndrome (SARS) in 2002 to 2003. The estimated value for R_0 indicated that the outbreak would lead to high cases of mortality which eventually turned out not to be the case. The deterministic SIR model which was used did not model the low transmission rates among quarantined individuals and also the high transmission rate in hospitals where contact rates are assumed to be high [8].

Another limitation is that the contact patterns are not realistically accounted for in the above deterministic modeling. The dynamics of a disease is subject to various important aspects such as age groups. Diseases spread differently in varying age groups and there are different incubation periods for people in different age group categories. It also depends on the type of contact involved between the susceptible and infected individuals for a successful transmission to occur, for example the contact between individuals is more intimate if they are sharing a home than at work place [8]. Deterministic models neither provide insight on the role of chance in the spread of a disease nor provide confidence intervals on results.

2.3 The case of Influenza Pandemic in San Francisco, 'Spanish flu'

For the purpose of this thesis, real data for the daily Influenza Pandemic cases recorded at the health department in the city of San Francisco was used. The data covers 63 days of the epidemic from September 23^{rd} to November 22^{nd}. Previous occurrence of the Influenza in the United States had facilitated the awareness of the Influenza existence to the health department. This lead to the data being recorded during the autumn wave in San Francisco to be thoroughly inspected and hence it is believed to be highly reliable [9].

During the years 1918 and 1919, the United States experienced the infamous influenza pandemic that lead to loss of many lives. Influenza was caused by a virus A(H1N1) and was commonly termed as the 'Spanish flu'. In San Francisco, the incidence left approximately 6.7% dead and had a high prevalence where 28310 cases of infection were reported [9].

The government implemented some interventions as control measures to reduce the burden of the rapidly spreading Influenza. Some of the measures as explained by [10] included;

i Educating and creating the awareness to the public through education campaigns

ii Isolating those who had already contacted the infection from the general public to prevent the infection from spreading

iii People were encouraged to wear face masks

iv Social gatherings were also prohibited since it was believed to be one way through which the influenza spread really fast

This widely known pandemic occurred in three waves. In particular, the city of San Francisco suffered during the autumn wave(second wave) that took place from September to November in 1918 and was considered to be the most devastating of the three waves [11].

In order to model the Influenza Pandemic, we use a simple SEIR epidemic Model as suggested by [9] and [10]. Figure 1 illustrates the model together with its components.

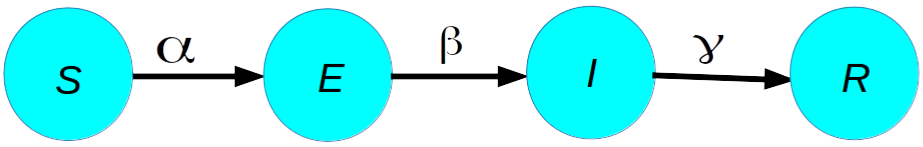


Figure 1. Compartmental diagram

The dynamics of the model in Figure 1 is further described by the following set of ordinary differential equations as follows

$$\begin{cases} \dfrac{dS}{dt} = -\alpha\, S\, I \\ \dfrac{dE}{dt} = \alpha\, S\, I - \beta\, E \\ \dfrac{dI}{dt} = \beta\, E - \gamma\, I \\ \dfrac{dR}{dt} = \gamma\, I \end{cases} \tag{2}$$

Where the state variables and the model parameters are described in Table 1 and Table 2 respectively.

Table 1. Description of parameters

Parameters	Description
α	Rate at which the susceptible humans become exposed
β	Progression rate from being exposed to infected
γ	The recovery rate

Table 2. Description of state variables

Variables	Description
S	The population of the susceptible
E	Exposed humans (latent stage)
I	Infected humans who can spread the infection.
R	The removed population either by death or due to recovery

Figure 2 represents the dynamics of the Influenza pandemic for the given 63 days. We used the data provided and specified the initial values for the ODE system as follows; number of initial infected was 4, both exposed and recovered populations was set to be 0 while the total population N was $28,310$. A model function was used to solve the system using ode45 in MATLAB, which is a built-in solver. The time starts from the point $t = 0$ which is a requirement for the ODE solver.

Numerical simulation indicates that the susceptible humans would decrease gradually. The number of individuals in the exposed class of humans starts to increase with decrease in the susceptibles and it reaches a peak. From there, a gradual decrease is observed. The infected class begins to grow shortly after the exposed class begins to grow. The short time difference here represents the incubation period in humans [9]. Then continues to increase until it reaches a peak from where the number of infected starts dropping. These infected individuals can now either recover, die naturally or die due to infection.

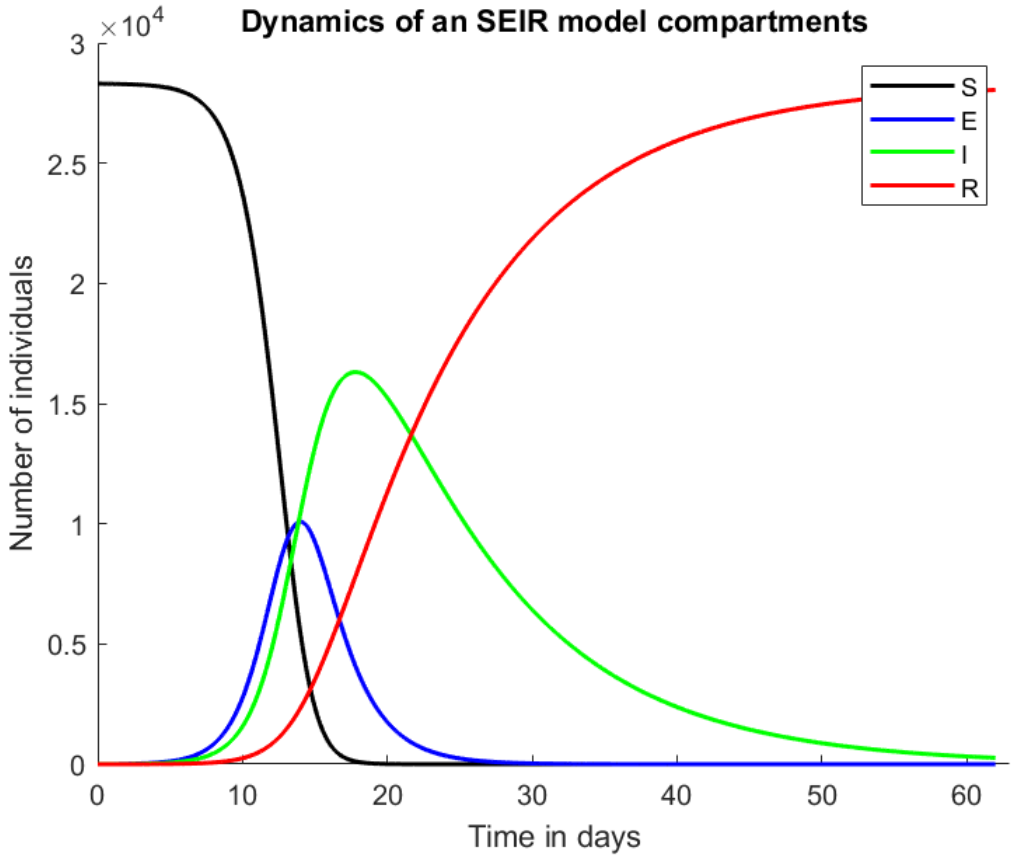


Figure 2. A graph showing the model compartments for the susceptible, exposed, infected and recovered.

Figure 3 shows a graph of how the data fits the model. We can observe that the model fits the data to some extent. The approach used to fit the data is the Least Squares approach that uses the *fminsearch* optimizer in MATLAB and the results from the fitting are plotted. Since we are dealing with ordinary differential equations which are non-linear, we can not use direct formulas that are used to compute LSQ estimate for liner models. We can only use numerical methods to do different approximations. The non-linear model to be minimized is given as

$$Y = f(X, \theta) + \epsilon, \tag{3}$$

where y represents the observed measurements, $f(X, \theta)$ is our model with X being the design matrix, θ is the unknown parameter whose value should be estimated and ϵ is the measurement error. In order to compute the least square estimate for the parameter θ, we minimize the sum of squares given in Equation 4 using numerical methods.

$$SS(\theta) = \sum_{i=1}^{n} [y_i - f(X, \theta)]^2. \tag{4}$$

$SS(\theta)$ is the sum of squared difference between the measurements and the model. Least squares approach gives us the parameter estimates that minimizes the sum of squares difference shown by Equation 4. This is done by an optimizer *fminsearch* in MATLAB. The LSQ approach mentioned here is from [12], where it has been described in details in both chapter 3 and 4.

The model given in Figure 1 has three parameters α, β and γ for which we needed to compute their estimates by fitting the influenza data. The parameter estimates obtained are given in Table 3 below.

Table 3. Parameter estimates

Parameters	estimated value
α	1.4979
β	0.5978
γ	0.7703

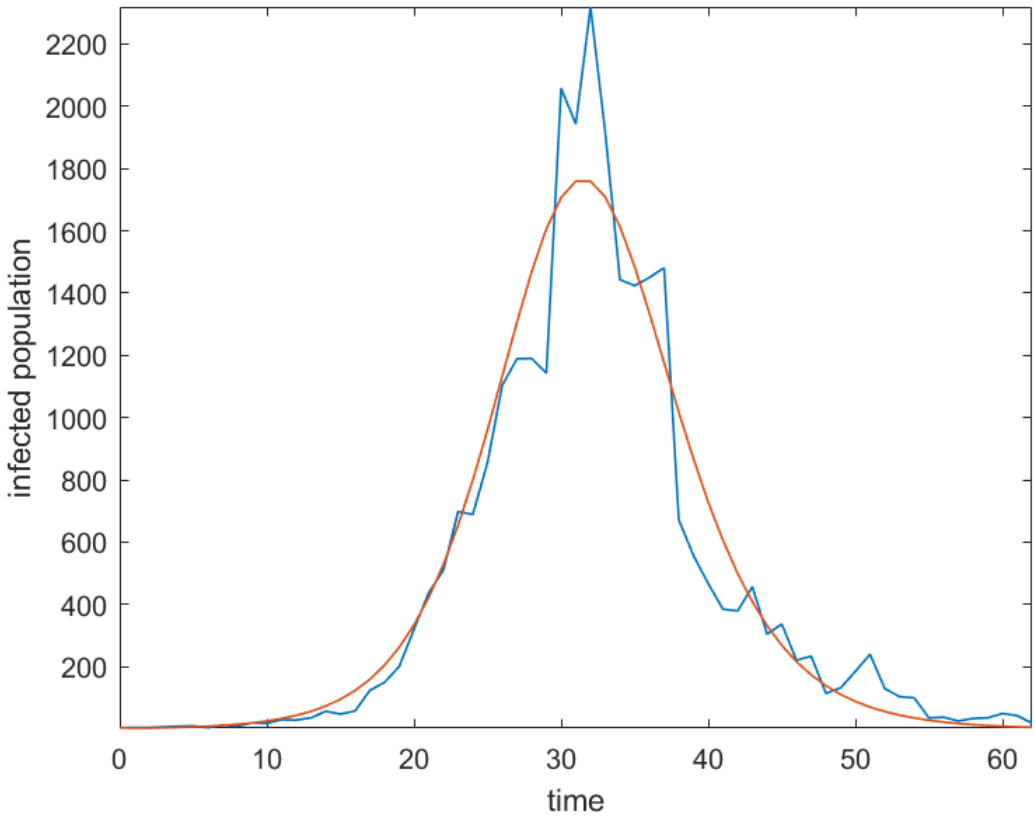


Figure 3. A graph of the model fitted with data

3 Discrete Modeling

3.1 Discrete Models

Discrete time stochastic models can be defined as "Markov chains approximations to the continous time models" [13]. A differential equation is said to be stochastic when some noise is added in an ODE. Therefore, discretization is a stochastic process in nature because of the randomness in the agents. For discrete models, a change in state variables only occurs at a countable number of intervals(points) in time.

A stochastic process is a family of random variables that depend on time. The time can either be discrete, where t = 0,1,2,..., or continuous, $t \in [0, \infty)$. Therefore the process can either be a "discrete-time" process such as random walk or a "continuous-time" process such as Brownian motion or the Wiener process [14].

For a particular set of input in discrete models, we obtain different variable output. The presence of randomness results into different results as the model output. For example, if we consider an independent variable x, if we have some model parameters associated with some uncertainty, then for the given x, different values of response variable y are generated each time.

The interval $[0, T]$ is discretized into $0 = t_0 < t_1 < ... < t_i < ... < t_T = T$ and the step size h_i is defined as

$$h_i = t_{i+1} - t_i$$

Discrete modeling is useful when considering small populations and when it is important to consider the randomness in the model parameter variation. However, discrete models are not cost effective since their computation requires multiple simulation runs when estimating the average outcomes of a particular epidemic [7]. Modeling interactions in discrete models depends on the susceptible individuals who acquire the infection though the interaction with the infected ones [8].

For all models we must note that when we have a given data for a particular disease outbreak the parameter estimates from that data should be equipped with some standard errors or some uncertainty. In addition, when we want to know if there is a chance that a disease would become extinct or would persist in the population in case of epidemic diseases, the question is well suited to be answered using stochastic modeling rather than deterministic modeling [15].

The dynamics of our SEIR discrete model can be described as follows; when an infected individual comes into a close contact with a randomly selected susceptible individual randomly in time, the susceptible one becomes exposed to the infection at a constant rate α. By close contact we mean a contact that would result into a new infected in a population given that, of the two individuals coming into contact, on is infected and the other one is susceptible. Otherwise the contact can not be considered effective [15].

The period of time that an individual spends in the exposed class is assumed to be independent and identically distributed with distribution F_E and mean $E[E] = \frac{1}{\beta}$. The infectious period has a distribution F_I with mean $E[E] = \frac{1}{\gamma}$. This conditions agrees with the deterministic model which assumes that the waiting times follows an exponential distribution.

The beginning of the epidemic is at time $t = 0$. With time, the epidemic is continuing to evolve and there is an addition of new infected individuals who recover eventually. This continues up to until first time T when there is no infected individuals in the population anymore. Therefore, there can be no further individuals becoming infected which implies that the epidemic will die out.

3.2 Gillespie Algorithm

In 1976, Dan Gillespie introduced the Gillespie algorithm. The algorithm appreciates some prominent steps from previous works done by Andrei Kolmogorov in 1931. Kolmogorov equation which is known today as the master equation was mainly based on differential equations corresponding to the time-evolution of stochastic processes that proceed by jumps. Later, other scientists such as William Feller contributed by developing a relationship between the already existing Kolmogorov's equations with random(stochastic) process [16].

Gillespie algorithm was inspired by the idea of the limitations of the regular deterministic modelling approach which proved to be inefficient in predicting cellular reactions for chemical and biochemical rate equations. The main difference between the algorithm and the old continuous deterministic way is that, it is a stochastic process in nature hence allowing a random simulation of the system with several reactants since each reaction process is explicitly simulated. As a result, Gillespie algorithm outputs a statistically correct feasible solution of a stochastic equation inform of a trajectory [16].

The implementation of Gillespie algorithm uses the number of possible reactions(events) and the rate at which they occur to generate randomly the time that next reaction will occur. Then uses probabilistic formulation to determine which reaction will occur next and are assumed to be random [17].

3.3 The Steps for the Gillespie Algorithm

The following are the steps for Gillespie Algorithm as explained by [18].

1. Initialize the population and the parameters.
2. Determine the reaction propensities $h_\mu c_\mu$ which is the probability of a reaction occurrence in a given time interval, where h_μ is the available number of combinations of the reactants, c_μ is the average probability that a certain combination h_μ reacts in the next time interval. Considering r reactions, we can say that the sum of the above propensities given by

$$a_0 \equiv \sum_{i=1}^{r} a_r$$

is the probability per unit time that a particular reaction will occur.

3. Generate two random numbers $rand_1$ and $rand_2$ which are uniformly distributed.

4. Using $rand_1$, compute for the time until the next reaction.
   For each change in time, the probability for a reaction to occur is considered to be a constant until the time a reaction will occur. This implies that the probability that no any reaction has occurred is an exponential decay

$$P_{unreacted} = e^{-a_0(\tau)}, \text{ where } \tau = t - t_{ref}.$$

If we consider a cumulative distribution for a reaction, we can say that

$$P_{reacted} = 1 - e^{-a_0(\tau)}. \tag{5}$$

We can differentiate Equation 5 with respect to time (t) to obtain

$$P(t) = \frac{dP_{reacted}}{dt} = a_0 e^{-a_0(\tau)}. \tag{6}$$

Equation 7 gives a distribution for the reaction times and can be written as

$$P(\tau) = a_0 e^{-a_0(\tau)}, \tag{7}$$

which is the time when the next reaction will occur.
Therefore, the reaction times that are being generated should be exponentially distributed. This can be done using the standard trick that converts $rand_1$ into an exponentially distributed function

$$\tau = \frac{1}{a_0} \ln(\frac{1}{rand_1}).$$

5. Determining the next reaction using $rand_2$.
   We assume two reactions R_1 and R_2. If R_1 has a propensity b_1 and R_2 a propensity b_2, then the probability that the next reaction to occur is R_1 is given by, $\frac{b_1}{b_1 + b_2}$. This concept can be generalized by saying that the probability that the next expected reaction is μ will be

$$P(\mu|\tau) = \frac{a_\mu}{a_0}.$$

   A random reaction is picked by selecting a random number $rand_2$ between 0 and 1 and figuring out where it lies in the reaction interval $\frac{a_1}{a_0}, \frac{a_2}{a_0}, ..., \frac{a_r}{a_0}$ that is in the interval $[0, 1]$. Therefore, reaction takes place next amounts to fining an integer μ for which the equation below is satisfied

$$\sum_{r=1}^{\mu-1} < rand_2 a_0 < \sum_{r=1}^{\mu}.$$

6. Update the time by adding τ to t, and also the state system.

7. Go back to step 2 until the target is achieved.

3.4 Discretization Approach (Euler method)

According to [8], the discrete time steps is given as follows:

1. Each susceptible individual in the population is drawn randomly from a uniform distribution. If the person drawn is infected, then he or she changes his state to exposed with probability α.
2. Each exposed individual changes his state to infectious state with probability β after the waiting time is elapsed.
3. Each infected individual moves from the infected class to the recovered class with probability γ after the waiting time is elapsed.
4. Each recovered individual remains resistant to the infection or removed from the population.

To implement step 1 we use a direct Euler method where the first ODE equation that represents the population of the susceptbles in the system of equations given in (2) is discretized as follows

$$\begin{aligned} \frac{dS}{dt} &= -\alpha SI, \\ \frac{S_{i+1} - S_i}{\Delta t} &= -\alpha S_i I_i. \end{aligned}$$

Doing some simple manipulations results into

$$\frac{S_{i+1}}{S_i} = 1 - \alpha I_i \Delta t,$$

where $\alpha I_i \Delta t$ fraction is bearing a negative sign to indicate that it is the fraction by which the individuals in the susceptible compartment decrease.

For the purpose of this work, we only have to discretize the differential equation that represents the compartment that consist of the susceptible individuals from the continuous model only. This is due to the reason that there is no latent time here and we can not tell how individuals in S get exposed to the infection. We only have the information that a person randomly gets infected and immediately moves to compartment E. However, after individuals move to E, we model this in a different way by just using the latent times It is a period of time that represents how long individuals spends in E before moving to I and also how long they stay in I before they move to R.

The main aim here is to double check if the latent time in compartments E and I are coming from an exponential distribution, then it should correspond to our system of ordinary differential equations. That is a limitation here because this ODE system compares or means that we have latent times but they only can be exponentially distributed. That is not the real case often since we can have indeed latent times which are minimum couple of days and only then something happens. So it is not always that this latent times would be exponential. The point in the influenza case is that we double check if by simply moving away from the classical modeling to latent time modeling by testing different latent time distribution we could get a better fit.

The other distributions considered include; weibull distribution, log-normal distribution and normal distribution.

4 Some Probability Distributions

4.1 Exponential Distribution

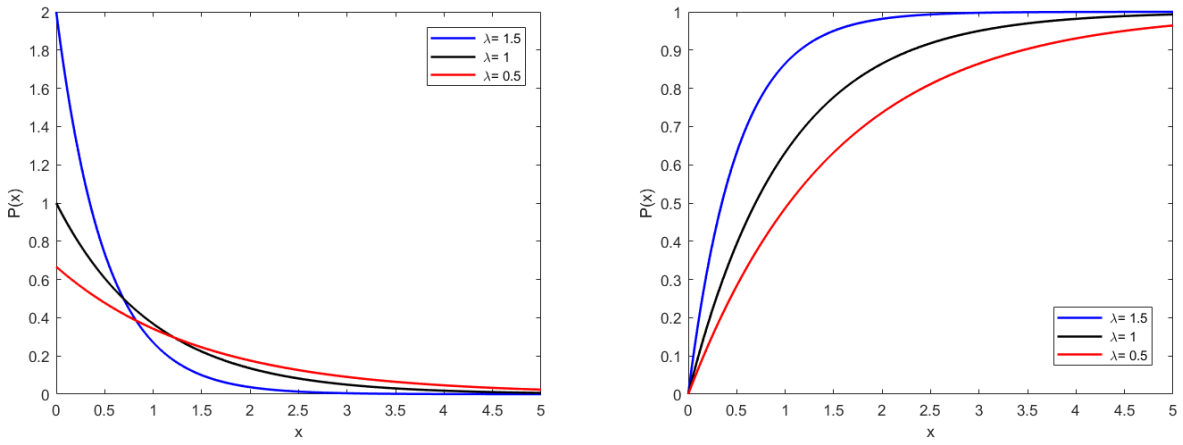
The exponential distribution is used to describe an event whose probability of such event occurring during a given period of time is proportional to the length of such time period. Thus, questions relating to the time needed for an event to occur is best modeled using an exponential distribution. The probability density of an exponential distribution with rate λ is given as

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

while the cumulative density function is given as

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

According to [19], the exponential distribution has a memoryless property. This implies that the probability that an event happens at a given period of time does not depend on how much time spent without the event happening.



(a) Probability density function for an exponential distribution(PDF)

(b) Cumulative density function for an exponential distribution(CDF)

Figure 4. From the diagram on the left hand side, the blue line shows that the exponential decay is high when the parameter value is large. Hence, as the rate increases, the mean value of the distribution increases.

4.2 Weibull Distribution

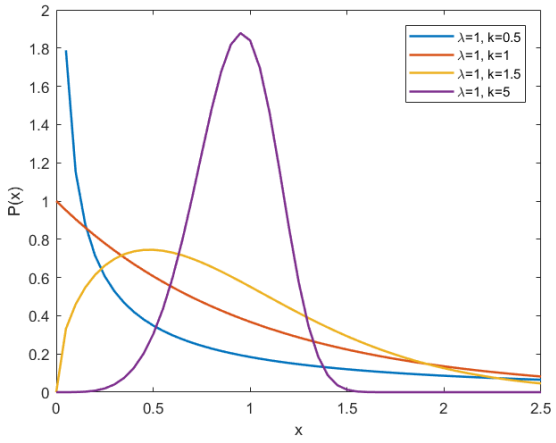
The probability density function(PDF) and cumulative distribution function of a random variable x which follows a weibull distribution is given by

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\frac{x}{\lambda}k\right) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

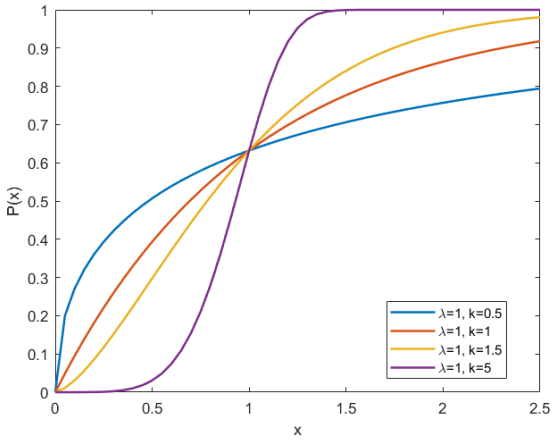
and

$$f(x;\lambda,k) = \begin{cases} 1 - \exp\left(-\frac{x}{\lambda}k\right) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

respectively [20, 21]. The parameters $k > 0$ and $\lambda > 0$ are the shape and scale parameters respectively.



(a) Probability density function for weibull distribution



(b) Cumulative density function for weibull distribution

Figure 5. From the figure on the left, when $k > 1$, as x approaches zero, the PDF tends to zero . Also, as $0 < k < 1$, the PDF tends to ∞ as x approaches zero from above. However, for $k = 1$, the PDF approximates to $\frac{1}{\lambda}$ as x approaches zero from above.

4.3 Normal Distribution

The normal distribution also known as the Gaussian distribution was initially described in the work of the German mathematician and physicist Johann Carl Friedrich Gauss in the year 1809. The probability density function (PDF) for normal distribution is given as

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{8}$$

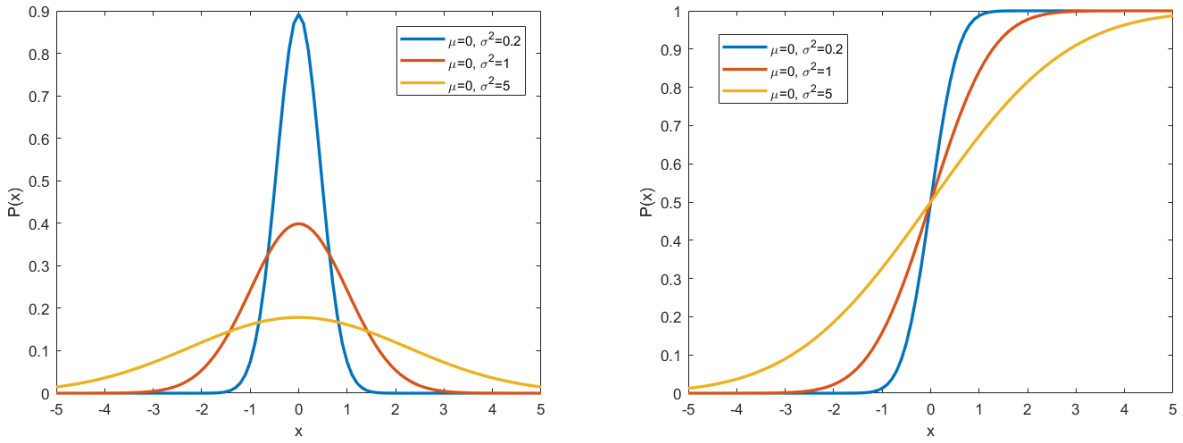
where μ is the mean or the expectation of the distribution, σ is the standard deviation and σ^2 is the variance. Thus, a normal distribution is always denoted by $N(\mu, \sigma^2)$. A special case of Equation 8 is when $\mu = 0$ and $\sigma^2 = 1$, which is referred to as the standard normal distribution often denoted by $N(0, 1)$. The PDF of a standard normal distribution is given as

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right),$$

while the cumulative distribution function (CDF) of the standard normal distribution often denoted by Φ is given as

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

Figure 6a and 6b shows the PDF and the CDF of a normal distribution.



(a) Probability density function for a normal distribution(PDF)

(b) Cumulative density function for normal distribution(CDF)

Figure 6. The red lines in both cases represents a special case of normal distribution called standard normal distribution which always has mean 0 variance 1. As the variance increases, the tail flattens at both ends of the distribution.

The normal distribution has the following characteristics [22, 23].

- The mean is the center point of the distribution
- The curve is symmetric in nature with mean=mode=median and also bell shape in nature
- The largest proportion of the values lies close to the center point, that is the mean. Hence, the further one goes from the mean, the fewer the scores.
- About 99.7% of the values lie within 3 times the standard deviation of the mean values.

4.3.1 Central Limit Theory and Law of Large Numbers

Given that $\bar{X}_n$ is the average of random variables X_i which are independent and identically-distributed (i.i.d) with theoretical mean and finite standard deviation σ. The law of large numbers states that as the sample size n increases, the probability that the average $\bar{X}_n$ is close to μ goes to unity. Furthermore, as the sample size n increases, the distribution of the average $\bar{X}_n$ approximates to the normal distribution with mean value and variance $\frac{\sigma^2}{n}$. That is, $N(\mu, \frac{\sigma^2}{n})$ [24].

5 Results From Numerical Simulations

Waiting times are the average times an individual spends in a compartment. Given our model Equation 2, we have two types of waiting times to consider. Average time that a human stays in the exposed class, $\frac{1}{\beta}$. Average time that a human stays in the infected class before recovery, $\frac{1}{\gamma}$. We use the distributions discussed in section 4 for the waiting times to do different numerical simulations in this chapter.

5.1 Discrete Case: Exponential Distribution for waiting times

Figure 7 illustrates how the data fits the model when the waiting times, $\frac{1}{\beta}$ and $\frac{1}{\gamma}$ are assumed to have been sampled from an exponential distribution and we used the command $exprnd$ from MATLAB. We fitted the data using the Least Squares method of optimization which gives us the best parameter estimate that can fit the data well. We did a time based simulation for our discrete function by taking 1000 simulation steps to do 50 repeated simulations. The whole population was 28310. At initial, we had the susceptible, $S_0 = 28306$, the exposed $E_0 = 0$, and the initial infected as $I_0 = 4$. The initial parameter values used for the optimizer were, $\alpha = 1.5030$, $\beta = 0.5934$, and $\gamma = 0.7712$. We can say that using exponential distribution, the model does not fit the data well.

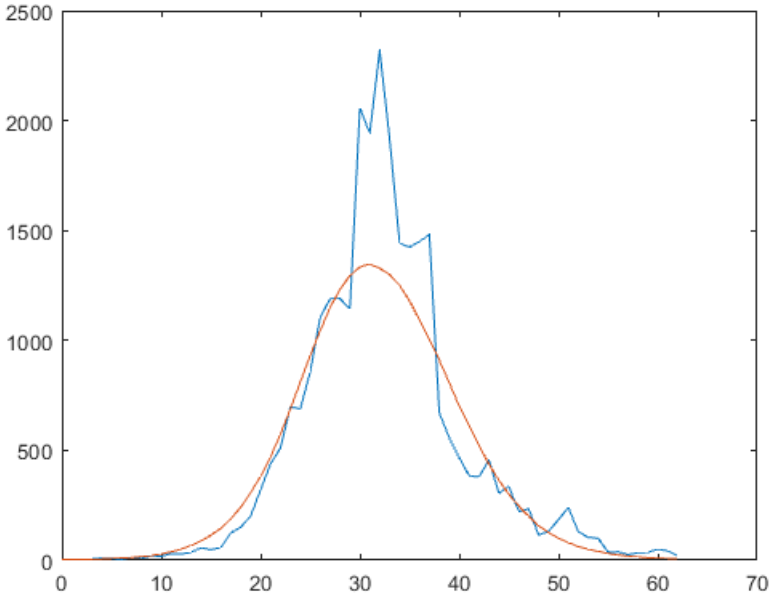


Figure 7. A graph of the model fitted with data using the exponential distribution for the waiting time. The blue line represents the data while the red line represents the model. The x and y axis represents the time and the cases of infection respectively.

Figure 8 shows the results of discrete approach using exponential distribution waiting times compared to the numerical solution. The total population is $28,310$. At initial point, the number of susceptible was taken to be $28,290$, the number of exposed individuals was equal to 0 , the total infected number was 20 and the recovered population was 0. The contact rate between the susceptible and infected was α 1.5030, the infection rate β was 0.5934 while the recovery rate γ was 0.7712.

In this case, 2,000 simulations steps were used in the run for 10 repeated simulation. The black, blue, magenta and red lines represents the numerical solution for the compartments of the SEIR model obtained from the MATLAB using $ode45solver$. The green lines represents the solution from discrete approach taking the waiting times to be exponentially distributed. We can observe that the discrete approach gives a similar solution to that of numerical one which is from a deterministic model when the waiting times are assumed to have been sampled from an exponential distribution. However, there are some visible fluctuations in the dynamics of susceptible and recovered individuals.

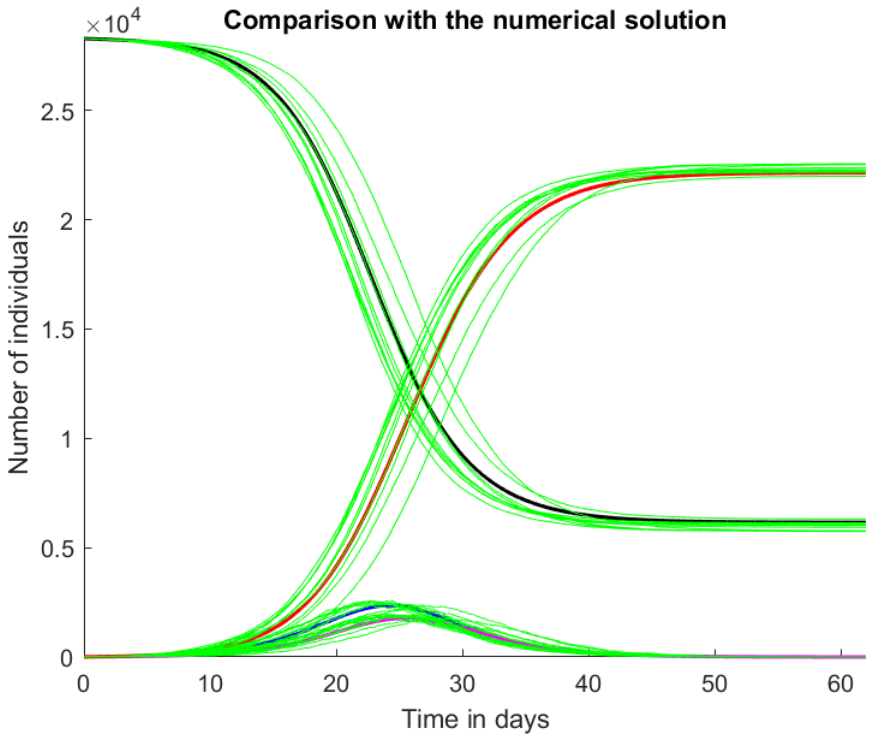


Figure 8. Discrete modeling with Exponential waiting times for 10 repeated simulations when $S_0 = 28,290, E_0 = 0, I_0 = 20, R_0 = 0$. The black, blue, magenta and red lines represents the dynamics of the Susceptible, Exposed, Infected and Recovered respectively obtained using $ode45$ solver while the green lines shows the solutions obtained from the discrete modeling using exponential distribution for waiting times.

5.2 Discrete Case: Weibull Distribution for waiting times

Figure 9 represents how the data fits the model when the waiting times, $\frac{1}{\beta}$ and $\frac{1}{\gamma}$ are assumed to have been sampled from a weibull distribution and we used the command $wblrnd$ from MATLAB. We fitted the data using the LSQ method which gives us the best parameter estimate that can fit the data well. We did a time based simulation for our discrete function by taking 1000 simulation steps to do 50 repeated simulations. The whole population was 28310. At initial, we had the susceptible, $S_0 = 28306$, the exposed $E_0 = 0$, and the initial infected as $I_0 = 4$. The initial parameter values used for the optimizer were, $\alpha = 1.5030$, $\beta = 0.4$, and $\gamma = 0.7712$, which are the model parameters. We also have other two additional parameter values that are as a result of weibull distribution which are $k = 1$ and $\lambda = 1$. They are the shape and scale parameters respectively. Having five parameters slows the operation of the optimizer and it ends up taking too long to give a solution. The optimizer has to go through all iterations to check and give the best

parameter estimate that could fit the data well for all five parameters at each iteration. The more the parameters to be estimated, the slower the optimizer becomes. We can say that using weibull distribution, the model fits the data but not quite well.

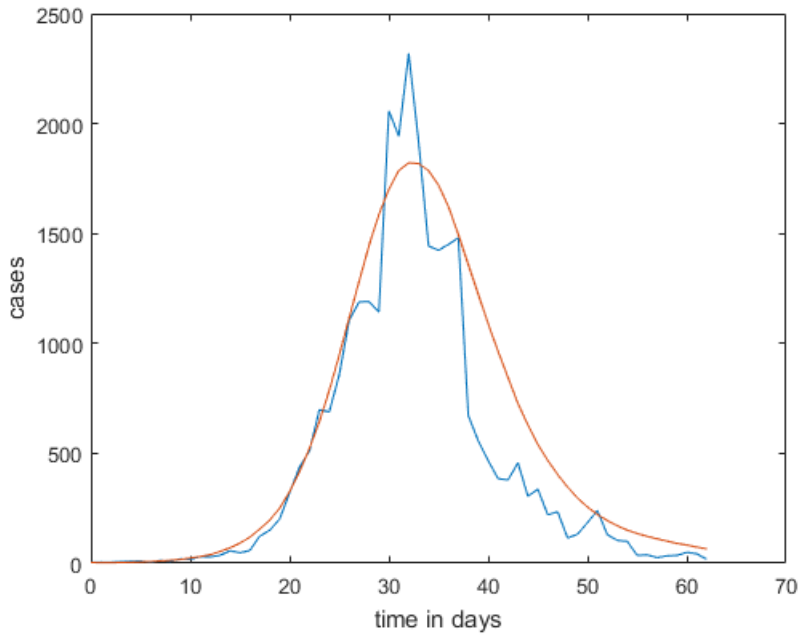


Figure 9. A graph of the model fitted with data using the weibull distribution for the waiting times. The blue line represents the data while the red line represents the model.

Figure 10 shows the results of discrete approach using weibull distribution waiting times compared to the numerical solution. The total population is $28,310$. At initial point, the number of susceptible was taken to be $28,290$, the number of exposed individuals was equal to 0 , the total infected number was 20 and the recovered population was 0. The contact rate between the susceptible and infected was α 1.5030, the infection rate β was 0.4 while the recovery rate γ was 0.7712. The shape parameter for webull distribution was $k = 1$ and the scale parameter was $\lambda = 1$

In this case, 2,000 simulations steps were used in the run for 10 repeated simulation. The black, blue, magenta and red lines represents the numerical solution for the compartments of the SEIR model obtained from the MATLAB using $ode45solver$. The green lines represents the solution from discrete approach taking the waiting times to be exponentially distributed. We can observe that the discrete approach gives a similar solution to that of numerical one which is from a deterministic model when the waiting times are assumed

to have been sampled from weibull distribution. However, there are some visible discrepancies in the dynamics of susceptible and recovered individuals. The exposed class seems to have almost fitted well with very minimal fluctuations. We can also observe that the infected class of individuals fits best here with the discrete solutions following the numerical one well for this compartment.

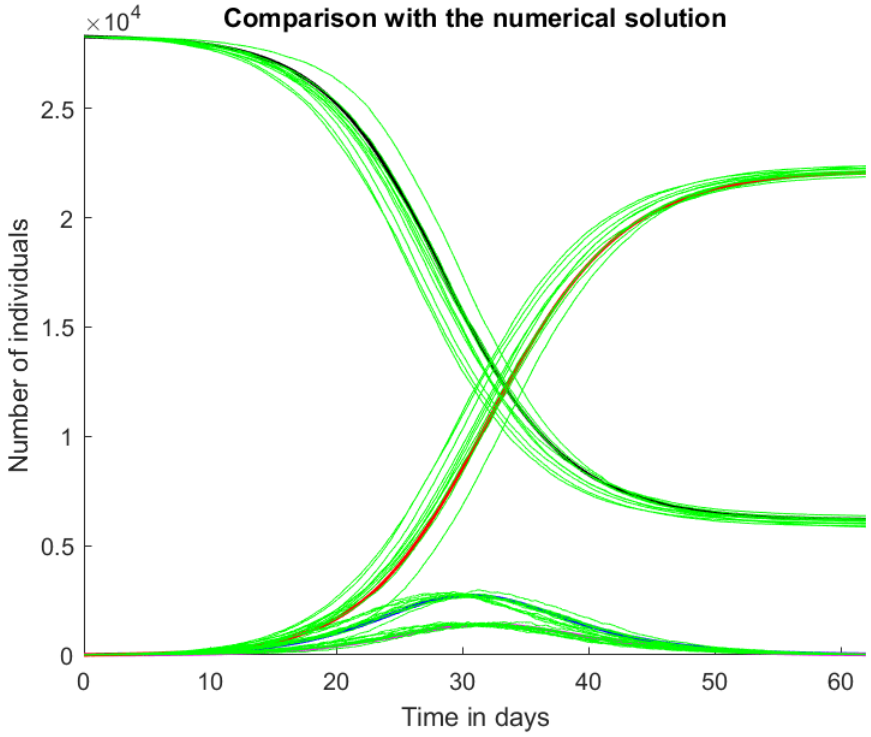


Figure 10. Discrete modeling with weibull waiting times for 10 repeated simulations when $S_0 = 28,290, E_0 = 0, I_0 = 20, R_0 = 0$. The black, blue, magenta and red lines represents the dynamics of the Susceptible, Exposed, Infected and Recovered respectively obtained using $ode45$ solver. The green lines shows the solutions obtained from the discrete modeling using weibull distribution for waiting times.

5.3 Discrete Case: Normal Distribution for waiting times

Figure 11 represents how the data fits the model when the waiting times, $\frac{1}{\beta}$ and $\frac{1}{\gamma}$ are assumed to have been sampled from a normal distribution and we used the command $randn$ from MATLAB. We fitted the data using the Least Squares method of optimization which gives us the best parameter estimate that can fit the data well. We did a time based simulation for our discrete function by taking 1000 simulation steps to do 50 repeated simulations. The whole population was 28310.

At initial, we had the susceptible, $S_0 = 28306$, the exposed $E_0 = 0$, and the initial infected as $I_0 = 4$. The initial parameter values used for the optimizer were, $\alpha = 1.5030$, $\beta = 0.4$, and $\gamma = 0.7712$. The probability density function for a normal distribution has two parameters mean and standard deviation, i.e $N(\mu, \sigma)$. In order to improve the efficiency of the code, we reduced the number of parameters by expressing the variance in terms of the mean parameter by, $\sigma = \frac{\mu}{3}$. This was done by using the idea that 99.7% of the values lie within 3 times the standard deviation of the mean values. From Figure 11, we can see that the normal distribution fits the data very well.

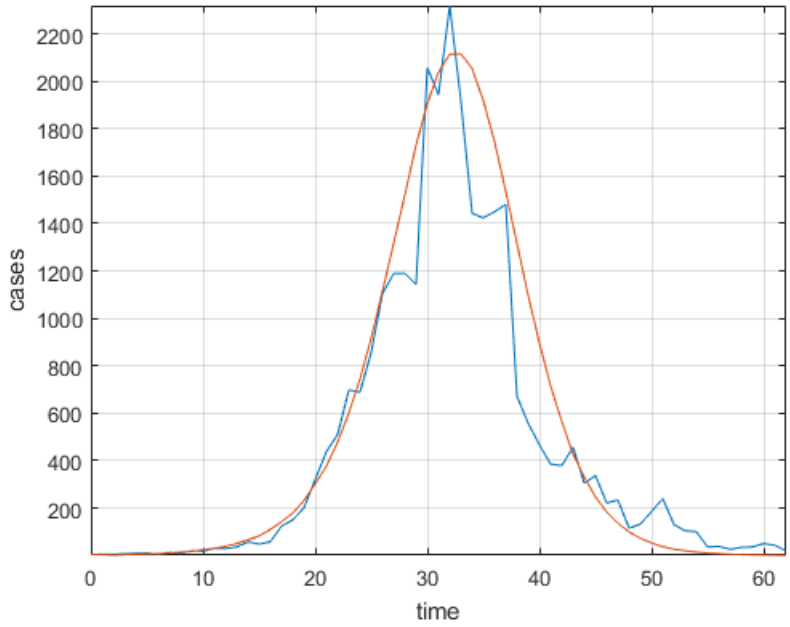


Figure 11. A graph of the model fitted with data using the normal distribution for the waiting times. The blue line represents the data while the red line represents the model.

Figure 12 shows the results of discrete approach using normal distribution waiting times compared to the numerical solution. The total population is $28,310$. At initial point, the number of susceptible was taken to be $28,290$, the number of exposed individuals was equal to 0 , the total infected number was 20 and the recovered population was 0. The contact rate between the susceptible and infected was α 1.5030, the infection rate β was 0.4 and the recovery rate γ was 0.7712.

In this case, $2,000$ simulations steps were used in the run for 10 repeated simulation. The black, blue, magenta and red lines represents the numerical solution for the compartments

of the SEIR model obtained from the MATLAB using $ode45solver$. The green lines represents the solution from discrete approach taking the waiting times to be normally distributed. We can observe that the discrete approach gives a similar solution to that of deterministic model when the waiting times are assumed to have been sampled from a normal distribution. The fluctuation between these two solutions are very minimal and the two seems to remain very close to each other despite the randomness that is known to be in the discrete case. The dynamics of the susceptible, exposed, infected and recovered seems to behave as expected.

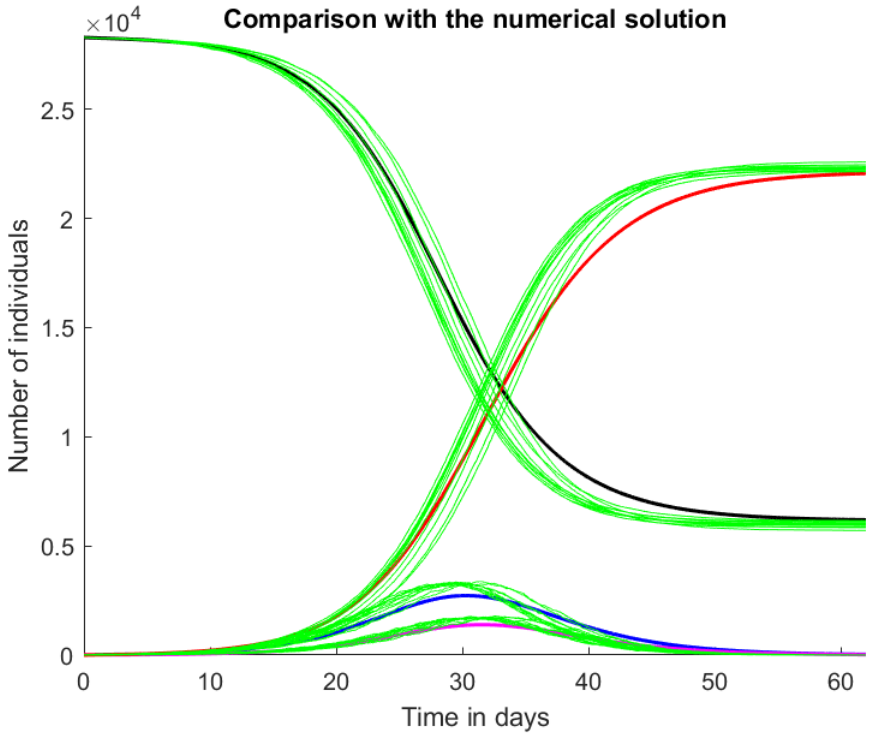


Figure 12. Discrete modeling with normal waiting times for 10 repeated simulations when $S_0 = 28,290, E_0 = 0, I_0 = 20, R_0 = 0$. The black, blue, magenta and red lines represents the dynamics of the Susceptible, Exposed, Infected and Recovered respectively obtained using $ode45$ solver. The green lines shows the solutions obtained from the discrete modeling using normal distribution for waiting times.

Figure 13 represents the results of discrete approach using normal distribution waiting times compared to the numerical solution. The total population is $28,310$. At initial point, the number of susceptible was taken to be $28,306$, the number of exposed individuals was equal to 0 , the total infected number was 4 and the recovered population was 0. The contact rate between the susceptible and infected was α 1.5030, the infection rate β

was 0.4 while the recovery rate γ was 0.7712.

In this case, 2,000 simulations steps were used in the run for 10 repeated simulation. The black, blue, magenta and red lines represents the numerical solution for the compartments of the SEIR model obtained from the MATLAB using $ode45solver$. The green lines represents the solution from discrete approach taking the waiting times to be normally distributed.

We can observe that when the amount of initial infected is decreased from 20 to 4, the system becomes unstable as the results in Figure 13 shows and the accuracy of the solution is not high. Also we can see that the fluctuation between the numerical solution and the discrete solution becomes higher when we lower the initial infected.

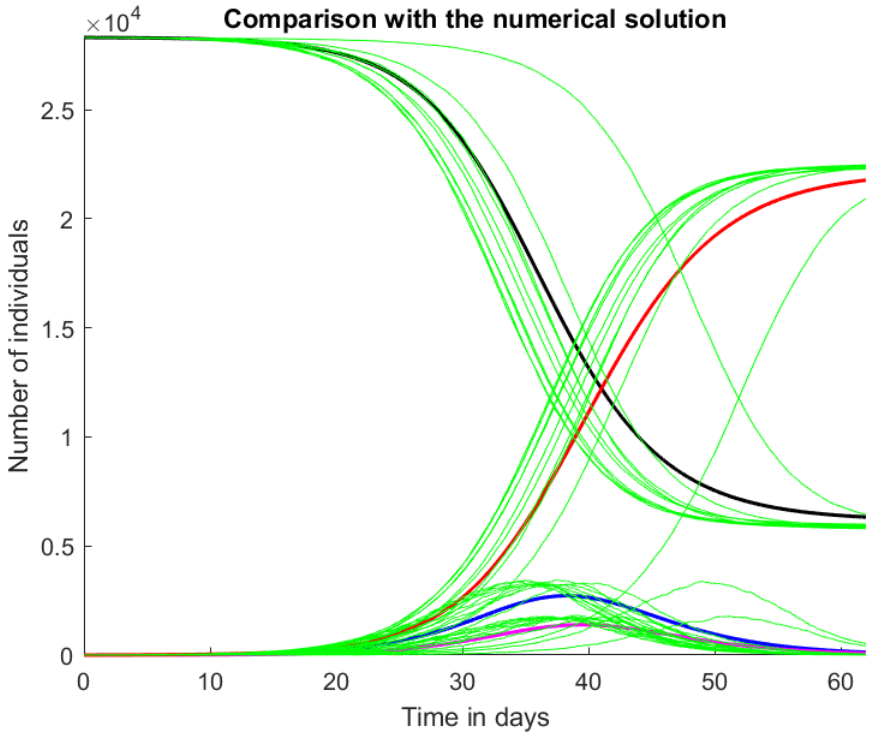


Figure 13. Discrete modeling with normal waiting times for 10 repeated simulations when $S_0 = 28,306, E_0 = 0, I_0 = 4, R_0 = 0$. The black, blue, magenta and red lines represents the dynamics of the Susceptible, Exposed, Infected and Recovered respectively obtained using $ode45$ solver. The green lines shows the solutions obtained from the discrete modeling using normal distribution for waiting times.

6 CONCLUSION

The goal of this study was to double check if the waiting times in an SEIR model comes from an exponential distribution using the 1918 San Francisco data of Influenza as a case study . Basic ODE modeling which is also known as Equation Based Modeling (EBM) only compares to exponential waiting times. This is a major limitation of this approach which lead us to use discrete case which is more efficient and also enables us to compare waiting times assumed to be sampled from different distributions. Hence we can see if we could get better fit for our data. The ODE system was transformed into a discrete model by Euler approach. From the numerical simulations, we have seen that there is a connection between the ODE and discrete modeling approach that they produce same results.

We performed numerical simulations of comparisons for cases of four different distributions for waiting times between the four classes of an SEIR model. We checked what kind of distributions we have for the unaccounted time which goes on in between from one class to another rather than having the classical way of just having waiting times to be coming from an exponential distribution. We considered other distributions which are weibull and normal and we compared their solutions to that of the ODE solver.

The normal distribution with mean $\mu = \frac{1}{\theta}$ and standard deviation $\sigma = \frac{\mu}{3}$ where θ represents the model parameters β and γ and the μ represents the waiting time in the exposed class E given as $\frac{1}{\beta}$ and the waiting time in the infected class I given as $\frac{1}{\gamma}$ gave a better fit for our data. This can be seen from the solution obtained from an optimizer in Figure 11 that shows that normal distribution gives a better fit for the data than the rest of the distributions.

Using different number of initial infected in the model, we can see the the variability increases when the amount of initial infected is lowered but remains low when the amount of initial infected is considered to be a bit higher. We need a enough number of initial infected for discrete modeling to give good results. Even with big populations, it is expected to get some uncertainties if initial number of infected is small. Comparing Figure 12 and Figure 13, with initial infected being 20 and 4 respectively, we can say that discrete modeling is sensitive to the initial values used. The system is mores stable at 20 as compared to at 4.

Discrete modeling allowed us to try different distributions hence we were able to get a more realistic fitting. Also it brings us more closer to reality since it is not always that a disease outbreak would become an epidemic. For example, if we have one initial infected in a big population, he/she may or may not infect other people. It can be that nothing happens as far as disease spread is concerned since the individual can go from class I to R before anybody gets infected. By the use of differential equation system in deterministic case, it is the expected value of what happens but in real life it may not always happen that way. Therefore discrete approach of modeling is considered more suitable in this study over the deterministic approach.

6.1 Future Work

For the extension of this thesis in future, we will implement the Gillespie algorithm following the steps outlined in Sub-section 3.3. Then compare the results from Gillespie algorithm to those of deterministic models. Also it would be important to compare Gillespie algorithm results to those of Euler approach to see which is best given that both methods are discrete in nature. In addition, we shall also try other distributions such as the Log-normal and Gamma distribution to see if they can be used to model the waiting times for this San Francisco 1918 Influenza data.

REFERENCES

[1] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[2] Jill M Anderson, Adrienne A Byrne, R Fields, L Segovia, and Randall J Swift. Some simple epidemic models, 2006.

[3] MG Garner and SA Hamilton. Principles of epidemiological modelling. *Revue Scientifique et Technique-OIE*, 30(2):407, 2011.

[4] Fred Brauer and Carlos Castillo-Chavez. Basic models in epidemiology. In *Ecological Time Series*, pages 410–447. Springer, 1995.

[5] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan AJ Metz. On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology*, 28(4):365–382, 1990.

[6] JM Heffernan, RJ Smith, and LM Wahl. Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface*, 2(4):281–293, 2005.

[7] Elisa F Long and Margaret L Brandeau. Or's next top model: decision models for infectious disease control. *Tutorials in Operations Research*, pages 123–138, 2009.

[8] Nedialko B Dimitrov and Lauren Ancel Meyers. Mathematical approaches to infectious disease prediction and control. *INFORMS tutorials in operations research*, 7:1–25, 2010.

[9] Gerardo Chowell, Hiroshi Nishiura, and Luis MA Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface*, 4(12):155–166, 2007.

[10] Martin CJ Bootsma and Neil M Ferguson. The effect of public health measures on the 1918 influenza pandemic in us cities. *Proceedings of the National Academy of Sciences*, 104(18):7588–7593, 2007.

[11] Thomas A Garrett. Pandemic economics: The 1918 influenza and its modern-day implications. *Federal Reserve Bank of St. Louis Review*, 90(March/April 2008), 2008.

[12] Marko Laine Antti Solonen, Heikki Haario. Statistical analysis in modeling. LUT lecture notes on Statistical Analysis in Modeling, 2018.

[13] Linda JS Allen and Amy M Burgin. Comparison of deterministic and stochastic sis and sir models in discrete time. *Mathematical biosciences*, 163(1):1–33, 2000.

[14] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2001.

[15] Tom Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

[16] Gillespie algorithm. https://en.wikipedia.org/wiki/Gillespie_algorithm, Accessed May 2018.

[17] David Karig. Introduction to stochastic simulation with the gillespie method. *Princeton, nd Web*, 18, 2015.

[18] The gillespie stochastic simulation algorithm. http://people.uleth.ca/~roussel/C4000foundations/slides/25stochsim.pdf, Accessed May 2018.

[19] Exponentialdistribution. https://www.statlect.com/probability-distributions/exponential-distribution, Accessed May 2018.

[20] PE Oguntunde, OS Balogun, HI Okagbue, and SA Bishop. The weibull-exponential distribution: Its properties and applications. *Journal of Applied Sciences*, 15(11):1305–1311, 2015.

[21] Debanshee Datta and D Datta. Comparison of weibull distribution and exponentiated weibull distribution based estimation of mean and variance of wind data. *Proc. Int. J. Energy Inf. Commun*, 4(4):1–11, 2013.

[22] Richard G Brereton. The chi squared and multinormal distributions. *Journal of Chemometrics*, 29(1):9–12, 2015.

[23] Douglas G Altman and J Martin Bland. Statistics notes: the normal distribution. *Bmj*, 310(6975):298, 1995.

[24] Central limit theorem and the law of large numbers. https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6b.pdf, Accessed May 2018.