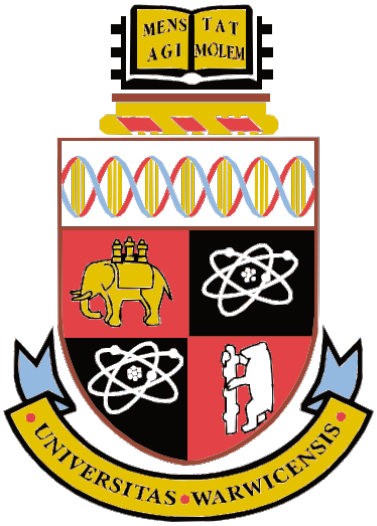


Persistence of Infectious Disease: understanding the role of infection characteristics using simulation and theory

by

Edward Gilbert

MA4K8 Scholarly Report

Submitted to The University of Warwick

Mathematics Institute

April, 2024

Contents

1 Introduction and Overview **1**

2 Context **1**
2.1 Why study persistence 1
2.2 So far on extinction 2

3 Methodology: The models **3**
3.1 The system of equations 3
3.2 Modelling stochasticity 5
3.3 Example model simulations 8

4 Methodology: The theory **10**
4.1 On steady state equilibrium 10
4.2 On natural resonance in our model 11
4.2.1 Eigenvalues of our system 11
4.2.2 The covariance matrix 12
4.2.3 Estimating time spent extinct with the Gaussian error function . . . 14

5 The results **16**
5.1 Looking at resonance in our model 16
5.1.1 The Fast Fourier Transform 16
5.1.2 Wavelets 17
5.2 Effect of changing population size and the amount of exposure/infection classes 19
5.3 Effect of changing average time infected 22
5.4 Effect of changing R_0 24

6 Discussion **28**
6.1 The results 28
6.2 Future work 29
6.2.1 Weaknesses of the Gaussian distribution approximation 29
6.2.2 Variations on the model 31
6.2.3 Programming the model 32
6.2.4 Testing the model 34

7 Conclusion **34**

A Further figures **38**

1 Introduction and Overview

In this dissertation I will be studying the persistence of infectious diseases by using simulation and theory. In the study of epidemiological models there is often a large focus on the more basic deterministic models. These models work well for large population sizes, however they fail to take into account the noisy behaviour observed when there are very few infected individuals. In this paper I will be focusing on this more noisy behaviour. I aim to discover how different epidemiological characteristics, such as the basic reproductive rate of the disease (R_0) and the distribution of infectious periods affect key metrics such as disease extinction lengths, as well as disease extinction rate. This has been achieved through the use of complex stochastic models, such as those explored by Dangerfield [11]. The decision to focus on persistence has been primarily driven by the study of Measles by Bolker [4], and a resulting desire to study persistence from a more detached position, than from having to modelling a real disease.

This report uses a theoretical approach, so although I discuss potential real world applications of these findings, the models have a limited basis in reality and should be treated as such. I am not trying to model any real world disease here, but instead seek to gain further understanding into the theory surrounding stochastic epidemic models and disease elimination. Parameters have been selected to have some basis in what one may typically expect from an infection of a real world population, however these decisions are not made with a specific disease in mind.

2 Context

In this section I am going to briefly outline some of the context around studying disease persistence before exploring persistence with my own models.

2.1 Why study persistence

Disease persistence is very important from a human perspective, as it is far more preferable to have a disease completely eliminated from a population as compared to continuing with a small number of cases. Disease control can be incredibly costly compared to elimination, especially when we're considering a large time horizon. If a disease is eliminated we no longer require a control program, which means costs such as vaccinations or other control measures are now eliminated. For example with Smallpox, developed countries are estimated to benefit by $350 million a year just from vaccination costs no longer being present post eradication [3].

It's difficult to overstate just how challenging disease elimination is, with humanity having only ever successfully eradicating two diseases: Smallpox and Rinderpest. For example,

the Guinea worm elimination program (GWEP) was started in 1986 with a target elimination date of 1995, then 2009, then 2015 and now 2030 [14]. We see these consistent push backs in elimination targets for diseases, due to the sheer difficulty of elimination. This challenge arises from not only trying to coordinate the use of control measures, but also from the fact that the dynamics of diseases tend to change greatly when we reach very low levels, making the effects of control measures harder to gauge [25].

Elimination and control can take vastly different amounts of work for a government to achieve and diseases can very easily be reintroduced to a population if a disease is not eradicated globally. This desire to ensure complete eradication over low level of a disease leads to some interesting questions mathematically.

2.2 So far on extinction

There has been a substantial amount of research done on disease persistence. In particular Bolker and Grenfell showed that the fade out proportion of a disease (the proportion of months with no cases) decreases as population rises for their model of Measles dynamics[4]. They showed differing behaviour for the SIR and SEIR models and also explored the use of imports, as well as age structure and periodicity. Conlan [9] expanded upon these results by looking at factors such as the proportion of epidemics terminated by fade outs of a disease.

Another similar area of study has been on time to extinction. Nasell [32] found that the time to extinction rose as population size rose for their SIR model. One idea explored by both Nasell as well as Bolker and Grenfell is the idea of critical community sizes, which is the size of the community necessary for a disease to persist. Nasell even explored the relationship between this and values of R_0, finding that higher R_0 values cause the critical community size to decrease for their model[1]. However neither of these papers directly explore the affects of changing parameters such as R_0 or the average infected period on persistence.

Despite all this, there has been limited research on more basic measures of persistence such as extinction rate and length. There is also limited study on the affect of changing parameters in models beyond the population size. This paper aims to look at how various key parameters affect these more basic measures, by looking at both a theoretical approach using various approximations, as well as by stochastic modelling.

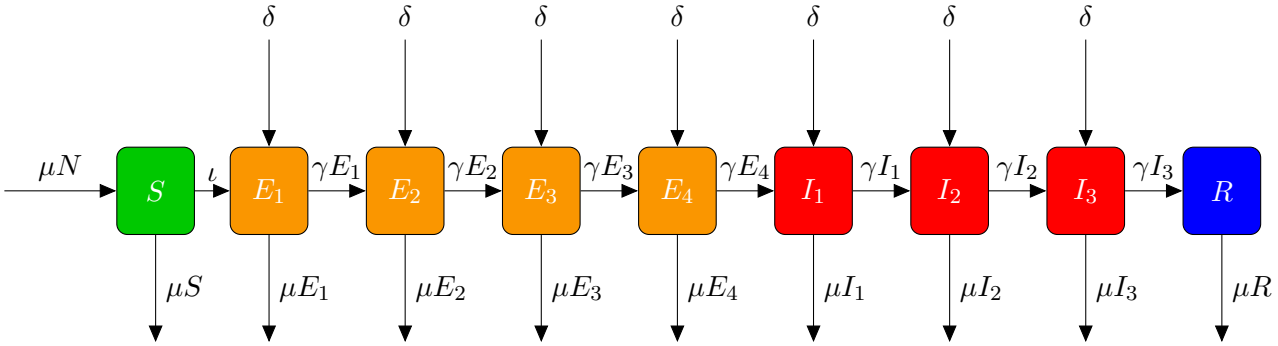
[1] I am specifically referring to the findings for the Martini model in section 2, as it is more realistic than the Bartlett model in section 3

3 Methodology: The models

3.1 The system of equations

Set of equations 1 is the system of equations used for most of the modelling. When any variation on this is used I will make it known

$$
\begin{aligned}
N &= S + \sum_{i=1}^{4} E_i + \sum_{i=1}^{3} I_i + R \\
\epsilon &= \frac{S}{N}\frac{\sqrt{N}}{5*365} \\
\frac{dS}{dt} &= \mu(N - S) - \beta\frac{S\sum_{i=1}^{3} I_i}{N} - \epsilon \\
\frac{dE_1}{dt} &= \beta\frac{S\sum_{i=1}^{3} I_i}{N} - \gamma E_1 - \mu E_1 + \frac{\epsilon}{7} \\
\frac{dE_2}{dt} &= \gamma(E_1 - E_2) - \mu E_2 + \frac{\epsilon}{7} \\
\frac{dE_3}{dt} &= \gamma(E_2 - E_3) - \mu E_3 + \frac{\epsilon}{7} \\
\frac{dE_4}{dt} &= \gamma(E_3 - E_4) - \mu E_4 + \frac{\epsilon}{7} \\
\frac{dI_1}{dt} &= \gamma(E_4 - I_1) - \mu I_1 + \frac{\epsilon}{7} \\
\frac{dI_2}{dt} &= \gamma(I_1 - I_2) - \mu I_2 + \frac{\epsilon}{7} \\
\frac{dI_3}{dt} &= \gamma(I_2 - I_3) - \mu I_3 + \frac{\epsilon}{7} \\
\frac{dR}{dt} &= \gamma I_3 - \mu R
\end{aligned}
\tag{1}
$$



$$\iota = \frac{\beta S(I_1+I_2+I_3)}{N}$$

$$\delta = \frac{S\sqrt{N}}{7(5*365)N}$$

Figure 1: SE^4I^3R compartmental diagram. This Figure shows how individuals travel through our system. Arrows represent where indivudals from each compartment will move to and the rates at which they do so

Where S represents number of susceptible individuals, R the number of recovered individuals, ϵ the number of imports[2], E_n the n^{th} exposure class and I_n the n^{th} infected class. In our model individuals get infected into the first exposure class, and then transition

[2]These imports are distributed evenly over our infectious and exposed classes

into subsequent classes at a constant rate γ. This transition process continues to move into the three infectious classes (where they can infect others) and then they finally recover. This use of multiple exposure and infectious classes may seem odd (and perhaps even unnecessary), however it can have a drastic effect on the dynamics of the disease. In the deterministic model the average time spent in each of the exposure/infectious classes is constant when we have one individual travelling through each class. However in our stochastic models each of these time periods will be exponentially distributed, with the time spent in multiple adjacent classes being Erlang distributed, which reflects reality to a greater extent [7]. The choice of 7 compartments was chosen such that when $\gamma = 1$, we get an average infected period of 1 week[3]. Next I'll explain each parameter.

γ is the rate of transition between each of the exposed/infectious classes. We assume our birth and death rates to be the same at μ and that all individuals are born susceptible. This means that without imports the population is constant. μ is our death rate, which we assume to be the same regardless of infection class. $\frac{1}{\gamma+\mu}$ is therefore the average time spent in each class. Therefore we can see that

$$\text{average infectious period} = \frac{\#\text{ classes which are infectious}}{\gamma + \mu}$$

It should be noted the vast majority of the time γ is many order of magnitude larger than μ, as by default I use $\gamma = 1$ and $\mu = \frac{1}{50\times365}$. Even with the simulations taking place over long time periods we can ignore the μ term here. This becomes noteworthy when calculating γ, as we can do it as just a function of average infectious period.

Another key parameter we'll be looking at is R_0, which is often thought of as describing the 'contagiousness' of a disease [12]. Four our model R_0 is defined as $\frac{\beta}{\text{average infectious period}}$

As we're studying persistence there is also a need for a low level of infectious imports ϵ in our model, to ensure the disease is never permanently extinct. The decision was made to make the amount of imports scale with $\sqrt{N}$. This was based on research conducted on Measles [13] which suggested that Measles imports scale with $\sqrt{N}$. These imports are representative of susceptible individuals who come into contact with infectious individuals from outside our population - hence we have S and $\sqrt{N}$ as scale factors. The scale factor of $\frac{1}{5*365}$ was chosen to keep imports relatively low to stop permanent extinction, while still allowing the disease to come back fairly frequently. I wanted there to be an equal chance of the imported disease being in any of the infectious classes, hence we have an import term in each infectious class and have a subsequent division by 7 - the number of infectious classes. [4]

[3]Please remember infected period $\neq$ infectious period

[4]This disease imports does cause the population to slowly increase over time. This will be a slight issue later when calculating frequencies of the number of infected, as our approach will rely on us having

Initially when constructing the model, I made the decision for the imports to come into the system in the form of new individuals. However, this contributes to a growth in the total population in the system over time, which will make some of the analysis further along more difficult. Therefore, I was inspired by the approach used by Alonso [1] where the susceptible individuals are gaining infection via imports [5]. As a result, I added the $\frac{S}{N}$ term. We will later see in the study that this level of imports ends up being rather small for various populations.

One final thing to note is that I opted to set the starting population (N_0) to 2^{14} (around 16000) by default. I used a multiple of 2 as many of my studies on population will use the powers of 2 for each population size. 2^{14} also represented the upper limit of what my laptop could simulate in a reasonable amount of time. I will touch on this more in section 6.2.

3.2 Modelling stochasticity

Although deterministic equations are useful for large populations, in reality we want to model this system stochastically, so we can explore disease persistence. In epidemiological modelling there are two common algorithms we use: The Gillespie algorithm [16] and the Tau Leap method [17]. These two algorithms dominate the literature on epidemiological modelling and for good reason [6] , as each of these methods is relatively simple and incredibly powerful.

With the Gillespie algorithm we start by creating a list of all the events that can occur in the system and the effect each event has on our population. We decide how long we want the simulation to run for (t_{max}) and then use the Gillespie algorithm as follows:

a fixed population. However, as scaling import with $\sqrt{N}$, for large enough N these imports will become insignificant

[5] In the real world this could be via travel to another population where the disease is still present and becoming infected

[6] There are plenty of other methods we can use for modelling the spread of infectious diseases. For example, the HKO method [21] (Page 16) can easily be adapted to epidemiological modelling. For very complex systems we can greatly increase computational efficiency with a clever selection of 'cells' for the algorithm, due to the use of subrules. However I opted not to use this algorithm as with the added complexity debugging would've been alot harder. It is also still far less quick than Tau Leaping. Similarly the adaptive Tau algorithm could have also been appropriate

Algorithm 1 Gillespie algorithm

1: Set $t = 0$
2: Calculate κ_i and κ for the initial stage of the system where κ_i is the propensities of each rule and κ is the sum of propensities of all rules
3: Set r_1 and r_2 as 2 independent uniformly distributed random numbers in the interval $[0, 1]$
4: Set $\tau = \frac{1}{\kappa} \ln r_1$
5: Set $t = t + \tau$
6: Determine which rule is to be executed at time t by finding m for which

$$\frac{1}{\kappa} \sum_{l=1}^{m-1} \kappa_l < r_2 \leq \frac{1}{\kappa} \sum_{l=1}^{m} \kappa_l$$

7: Execute rule m
8: Update propensities κ_i and κ
9: **if** $t < t_{max}$ **then**
10: Return to step 2
11: **end if**

We note that this method implies the time step between each reaction is an exponentially distributed random variable. For this to be the case, we must assume that the system is a Markov chain and that the systems next state is entirely dependent on the current state and not any of the states before. This is often colloquially characterised as the system having 'no memory'.

The Gillespie algorithm is an exact algorithm and is therefore very accurate to how we typically picture stochastic systems. An exact algorithm is one which is entirely event driven [8] and thus follows the exact dynamics given in the deterministic equations. Therefore over many runs we expect the average of the paths followed on the Gillespie algorithm to match that followed by our deterministic method. On the other hand exact algorithms can be very computationally intensive, so we often use methods such as the Tau Leap algorithm instead. With Tau Leap, we set a value τ as our timestep, and at each timestep we calculate the rates for each event i. For each possible event we then calculate the number $Num_i = Poisson(Rate_i * \tau)$. If executing the event i Num_i times results in a population where none of our population classes are below 0, we then execute the event Num_i times. Otherwise we execute the event the maximum amount of times possible such that none of our population classes go below 0. We then move onto the next event, until we've gone through all events.

Tau Leap is an approximation of Gillespie and is thus less accurate. However in this study I have decided to opt for Tau Leap due to computational efficiency. In Figure 2 one can see how long each method takes to run a simulation of 10 years for various population sizes with $\tau = 1$ over an average of 10 runs. For lower values of $\tau = K$, one would expect

the simulation to run for $\frac{1}{K}$ times longer.

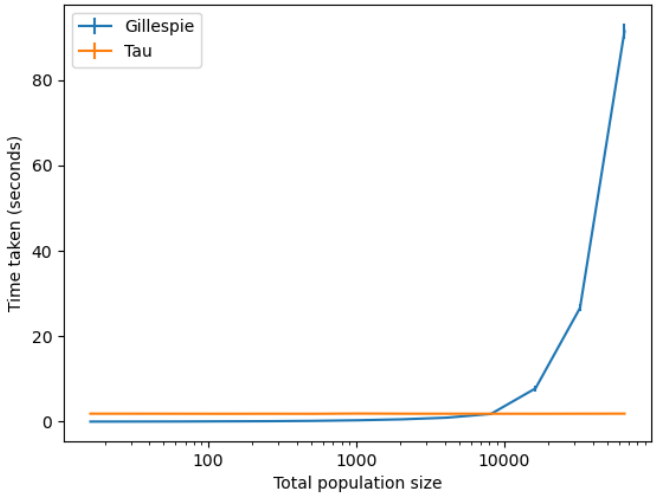


Figure 2: average time taken for each method at varying population sizes, with 98% confidence interval error bars included. Note the x-axis has a log 10 scale. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \gamma = 1, \tau = 1, R_0 = 5$

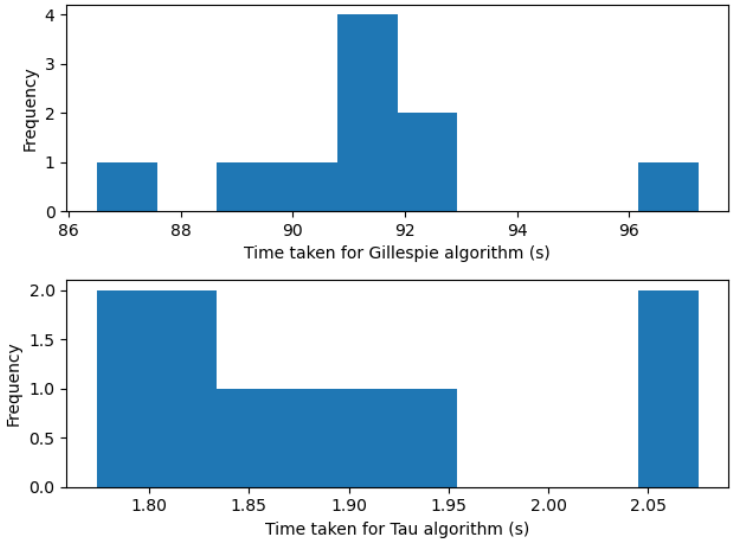


Figure 3: histograms of time taken for each method at $N_0 = 2^{18}$. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \gamma = 1, \tau = 1, N_0 = 2^{18} \approx 250,000, R_0 = 5$

The Tau Leap method is more computationally expensive for smaller populations, as seen in Figure 2. This is because after every time step we have to recalculate the rates. This is why the time the Tau Leap algorithm takes is largely independent of population size - the vast majority of the computational effort is spent calculating the rates, whereas there is minimal computational effort spent on applying the population changes, as they occur so

infrequently. The effect of this can be seen on Figure 3, with Tau Leap method showing far less variation. For a small population size the rates are also going to be small, therefore the timestep for the Gillespie algorithm will likely be much larger than a day, meaning Gillespie is far more efficient. However there eventually reaches a population size (in this case around 9000 individuals) where the time steps in Gillespie are far smaller than τ and we start to see Tau Leap become the more efficient choice.

One may question the use of $\tau = 1$, as this is fairly large, but it does mimmick the typical diurnal cycle we see for human behaviour. However this isn't without its problems. With the Tau Leap method, as well as the system of equations being used, one can only move through one class at a time. If γ is too high (in particular greater than one), this can cause serious issues. For example, if $\gamma = 2$, we would expect on average a person to move through 2 of the infected classes per day. However, if we have $\tau = 1$, a person will only be able to move through 1 infected class a day, meaning the simulation does not reflect our deterministic equations. Even with $\gamma = 1$ and $\tau = 1$, we will have some of this effect. At every time step we use a randomly Poisson distributed number to decide if a process goes ahead. However with $\tau = 1$ at the least it will take an individual one day to move from one state to another, but obviously there is no upper bound on how long it can take. This effect skews the average time infected and the average infectious period in our simulation higher than what we'd expect from using the Gillespie algorithm, or from our deterministic set of equations. In light of this, when we are looking at changing γ directly or indirectly by changing average infectious period, I shall use $\tau = \frac{1}{7}$ to prevent these issues. This value has been chosen as it allows some individuals to go through the all the infected classes in one day for a value of γ that is particularly high or for an average infectious period that's particularly short. However this will increase the time taken to run simulations by a factor of 7.

3.3 Example model simulations

Figure 4 has been produced using the Tau Leap method and show typical model results when $R_0 = 5$ over 40 years.

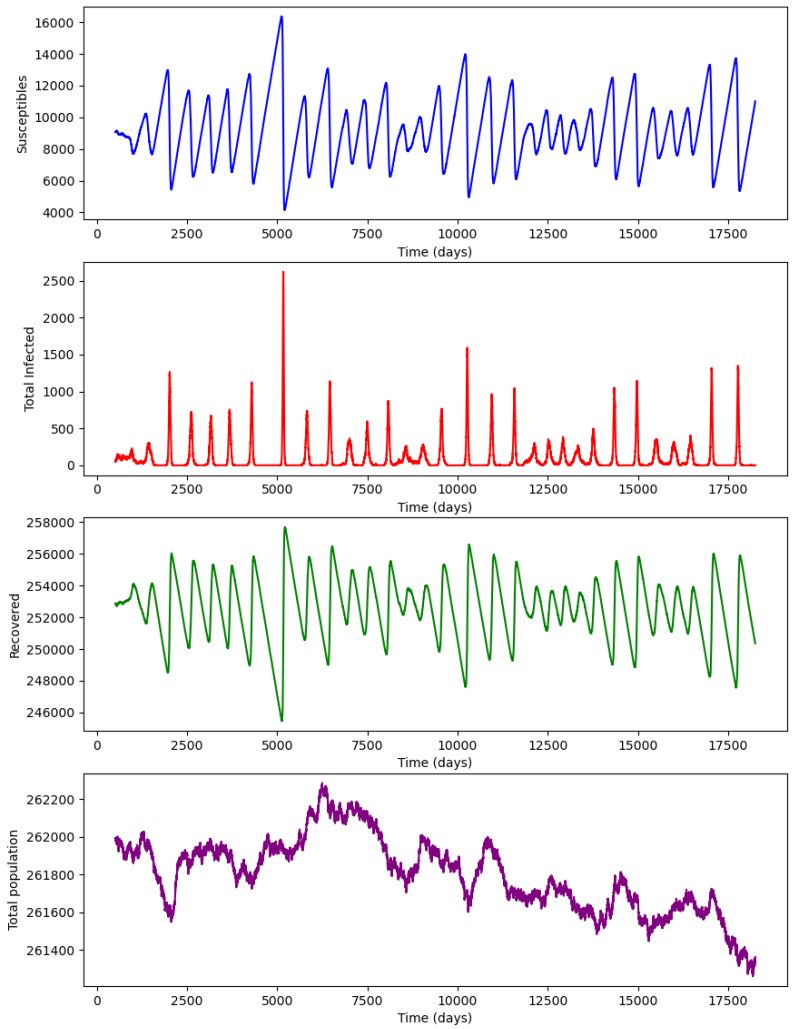


Figure 4: Plots for S, I, R & N classes over 40 years. This plot shows the stochastic nature of the dynamics. In particular one can see epidemic outbreaks, which will not occur in the deterministic model, where we would have a stable endemic level. $\mu = \frac{1}{50*365}, \gamma = 1, \tau = \frac{1}{7}, N_0 = 2^{18} =, R_0 = 5$

In Figure 4 we can see behaviour that appears wavelike - with variations in the total number of infected individuals appearing to occur around a somewhat stable equilibrium number. Our total population will vary somewhat due to the stochastic nature of the model, but ultimately stays around N_0.

We can even go a step further than this and see this oscillation in action on an SI plot,

such as Figure 5, where each dot represents a point in time, and one can see how over a 40 year time period we tend to see an orbit around some fixed point. Sometimes we observe a very large orbit around this point with some clear outer trails of points being visible. We can see where the disease becomes extinct and travels along the x axis as births occur, before gaining an import and thus rising up and making a loop around our stationary points S^* and I^*. As shown in the next section this ceroborates our theoretical approach.

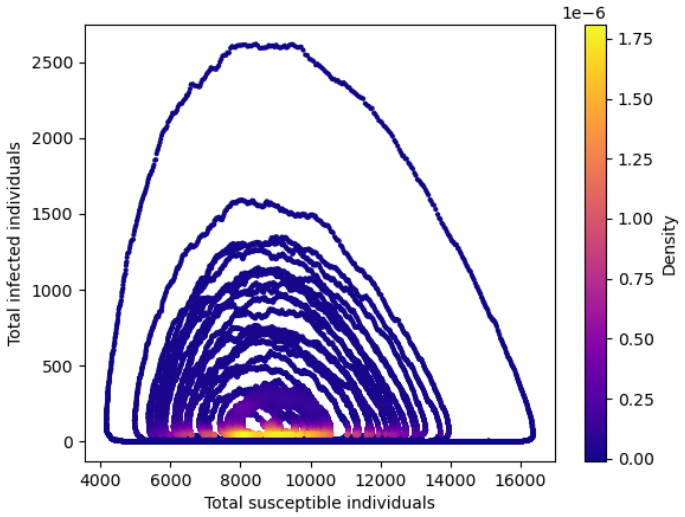


Figure 5: scatter plot for S and I where each time step is plotted, showing the stochastic dynamics in the phase space. Areas with a higher density of points are shaded yellow. $\mu = \frac{1}{50*365}, \gamma = 1, \tau = 1, N_0 = 2^{18}, R_0 = 5$

4 Methodology: The theory

4.1 On steady state equilibrium

We can see from Figure 5 that our stochastic simulations still appear to fluctuate around a steady state. One obvious question is what does theory suggest this steady state is?

The most obvious way to do this would be to look first at the deterministic equations and to find the non-trivial equilibrium solution. Analytically we can easily see that:

$$R^* = \frac{\gamma}{\mu} I_3^*$$

However the presence of imports and the fact they scale with our population greatly complicates the process of finding an equilibrium in understandable terms.

In fact, due to all the different infectious states we have, as well as the $\sqrt{N}$ term, any algebraic solutions we get will be rather cumbersome and won't provide much insight.

One easy approach will thus be to simplify the model slightly, while still keeping the same mechanical process. To do this we 'reduce' our system down to an SIR system.

I will use the parameters from the original model 1, but adapt these for the simplified SIR approach. We know for $\frac{4}{7}$ of our infected states the infection is not transmitted - so if in our simplified model everyone in class I is equally infectious then we must scale β appropriately. We also know that our average time in all the infected classes $\approx \frac{7}{\gamma}$. As a result of this we know our system will behave (somewhat) similar to the equations below:

$$\begin{aligned} N &= S + I + R \\ \frac{dS}{dt} &= \mu(N - S) - \frac{3}{7}\beta\frac{SI}{N} \\ \frac{dI}{dt} &= \frac{3}{7}\beta\frac{SI}{N} - \frac{\gamma}{7}I - \mu I \\ \frac{dR}{dt} &= \frac{\gamma}{7}I - \mu R \end{aligned}$$

From this system we can find a set of non-trivial equilibria in terms of N:

$$\begin{aligned} S^* &= \frac{1}{R_0}N \\ I^* &= \frac{7\mu}{\gamma+7\mu}\left(1 - \frac{1}{R_0}\right)N \\ R^* &= \frac{\gamma}{\gamma+7\mu}\left(1 - \frac{1}{R_0}\right)N \end{aligned}$$

For this model our value of R_0 end up being $\frac{3\beta}{\gamma+7\mu}$

To put things in perspective for $\mu = \frac{1}{50*365}$, $\gamma = 1$ and $R_0 = 5$, as proportions of our total population, we end up getting $S^* = 0.2$, $I^* \approx 0.0003$ and $R^* = 0.7997$. This aligns very well with what we see in Figure 4, where we often have very low levels of infectious individuals in our population relative to the amount of susceptible and recovered individuals.

4.2 On natural resonance in our model

4.2.1 Eigenvalues of our system

Our model contains a large amount of stochastic aspects and tends to feature periodicity, as partly illustrated by Figure 5. There is a simple way of studying the periodicity present in our models and that is by looking at the eigenvalues of the deterministic equations, as touched on in the Population dynamics module. To do this I am taking influence from [22] page 42. Similar to how we've found the covariance matrix by simplifying our deterministic equations down to 3 classes (the SIR model), we can simplify our deterministic equations down to 4 classes (the SEIR model) and then find the eigenvalues of the system. Using 4 classes will make our results more realistic, whilst not making our characteristic polynomial

too complicated. The appropriate SEIR system is shown below:

$$\begin{aligned} N &= S + E + I + R \\ \frac{dS}{dt} &= \mu(N - S) - \beta\frac{SI}{N} \\ \frac{dE}{dt} &= \beta\frac{SI}{N} - \frac{\gamma}{4}E - \mu E \\ \frac{dI}{dt} &= \frac{\gamma}{4}E - \frac{\gamma}{3}I - \mu I \\ \frac{dR}{dt} &= \frac{\gamma}{3}I - \mu R \end{aligned}$$

Note in this case we do not need to scale β like we did in the 3 classes case, as we now have separate infectious and exposed classes again. We want to try to condense each of our exposed classes into one class, and we do this by noting that in our original model the average time spent in the exposed classes would be $\approx \frac{4}{\gamma}$, so we use $\frac{\gamma}{4}$ as the rate at which exposed individuals become infected. We use a similar logic for our recovery rate. If we set $N = 1$ and note the recovered class is thus redundant, this overall set of equations gives us the following characteristic polynomial:

$$\mu(R_0 - 1)\left(\mu + \frac{\gamma}{4}\right)\left(\mu + \frac{\gamma}{3}\right)x^3 + \mu R_0\left(2\mu + \frac{\gamma}{4} + \frac{\gamma}{3}\right)x^2 + \left(\mu R_0 + 2\mu + \frac{\gamma}{4} + \frac{\gamma}{3}\right)x + 1 = 0$$

The roots from this polynomial will contain a complex conjugate pair, where the imaginary part ω corresponds to the period of oscillation T in our model, with $T = \frac{2\pi}{\omega}$. For our usual parameters $\mu = \frac{1}{50*365}, \gamma = 1$ and $R_0 = 5$, we find that $\omega \approx 0.0056$, which gives us a period of approximately 1200.

We must consider that this period has been calculated for an approximation for our system. Therefore stochastic simulations may not show exactly the same period.

4.2.2 The covariance matrix

I will also look at the more theoretic based approaches discussed in the appendix of [11], particularly the use of the covariance matrix. This requires use of diffusion approximations[27] [26], which involves looking at an equivalent diffusion process instead of our stochastic process[19]. The diffusion approximation is incredibly useful for us, as it allows us to think about the process as diffusion away from deterministic fixed points.

The covariance matrix σ^2 is defined by the set of equations

$$B\sigma^2 + \sigma^2 B^T + G = 0$$

where:

$$\begin{aligned} G &= \begin{pmatrix} \sum_e (\Delta s_e)^2 f_e & \sum_e (\Delta s_e)(\Delta i_e) f_e \\ \sum_e (\Delta i_e)(\Delta s_e) f_e & \sum_e (\Delta i_e)^2 f_e \end{pmatrix}\Bigg|_{s^*,i^*} \\ F &= \begin{pmatrix} \sum_e (\Delta s_e) f_e \\ \sum_e (\Delta i_e) f_e \end{pmatrix} \\ B &= \nabla F|_{s^*,i^*} \\ \sigma^2 &= \begin{pmatrix} \mathrm{Var}(s^*) & \mathrm{Covar}(i^*, s^*) \\ \mathrm{Covar}(s^*, i^*) & \mathrm{Var}(i^*) \end{pmatrix} \end{aligned}$$

Where e represents each event, Δx_e represents the change in x when event e takes place. f_e is the rate at which event e takes place. There are only 5 events in our SIR model that affect susceptible and infected individuals (birth, death of susceptible, death of infected, infection and recovery), therefore finding G turns out to be more simple than it appears, with each sum only having a maximum of 3 events to consider. Note when thinking about the covariance matrix, we are thinking about our population in terms of proportions, so in terms of $s^* = \frac{S^*}{N}$ and $i^* = \frac{I*}{N}$. We have to be very careful with this, as otherwise it is easy for errors to occur. Putting the above expression in terms of our equilibrium points, we get:

$$\begin{aligned} G &= \tfrac{1}{N^2} \begin{pmatrix} \frac{\frac{3\beta}{7} s^* i^*}{1} + \mu N + \mu s^* & \frac{-\frac{3\beta}{7} s^* i^*}{N} \\ \frac{-\frac{3\beta}{7} s^* i^*}{N} & \frac{\frac{3\beta}{7} s^* i^*}{N} + \frac{\gamma}{7} i^* + \mu i^* \end{pmatrix} \\ F &= \tfrac{1}{N} \begin{pmatrix} -\frac{\frac{3\beta}{7} s^* i^*}{1} + \mu N - \mu s^* \\ \frac{\frac{3\beta}{7} s^* i^*}{N} - \frac{\gamma}{7} i^* - \mu i^* \end{pmatrix} \\ B &= \begin{pmatrix} \frac{-\frac{3\beta}{7} I^*}{N} - \mu & \frac{-\frac{3\beta}{7} S^*}{N} \\ \frac{\frac{3\beta}{7} I^*}{N} & -\frac{\gamma}{7} - \mu + \frac{3\beta}{7} s^* \end{pmatrix} \end{aligned}$$

Note $B_{2,2}$ is the same as our expression for $\frac{di}{dt}$, all divided by i. As B is evaluated at the steady state, $\frac{di^*}{dt} = 0$, so $B_{2,2}$ is equal to 0. We then substitute in our values for s^* and i^*, giving us

$$\begin{aligned} G &= \tfrac{1}{N^2} \begin{pmatrix} 2\mu & -\mu\left(1 - \frac{1}{R_0}\right) \\ -\mu\left(1 - \frac{1}{R_0}\right) & 2\mu\left(1 - \frac{1}{R_0}\right) \end{pmatrix} \\ B &= \tfrac{1}{N} \begin{pmatrix} -\mu R_0 & \mu(R_0 - 1) \\ -\frac{+7\mu}{7} & 0 \end{pmatrix} \end{aligned}$$

We also note $\left(\sigma^2\right)_{1,2} = \left(\sigma^2\right)_{2,1}$, meaning our matrix σ^2 is symmetric, which makes finding σ^2 easier. We can then substitute in $\sigma^2 = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ to give us 4 sets of simultaneous equations [7]. From here one can easily find b, as it appears in one equation with neither a

[7] One of which will be redundant

nor c. From here we can substitute to find a and then c. Through some simplification we end up with the following:

$$\sigma^2 = \frac{1}{N}\begin{pmatrix} R_0 + \frac{\gamma}{7\mu} & \frac{-1}{R_0} \\ \frac{-1}{R_0} & R_0 - 1 + \frac{\mu(1+R_0)}{{R_0}^2\left(\frac{\gamma}{7}+\mu\right)} \end{pmatrix}$$

We can see that our covariance between s and i, $\frac{-1}{NR_0}$ is negative, indicating our values for s and i have an inverse relationship. This makes sense, as when i is high we expect s to be lower as individuals in the susceptible class have been infected. Interestingly we see that as R_0 gets larger this covariance becomes smaller, indicating that the relationship between the two becomes weaker.

Our variance for each element is represented on the 2 diagonal entries, and we have for s, $\frac{R_0}{N} + \frac{\gamma}{7\mu N}$ and for i, $\frac{R_0-1}{N} + \frac{\mu(1+R_0)}{N{R_0}^2\left(\frac{\gamma}{7}+\mu\right)}$. Both clearly get smaller as N gets larger. We also note that μ makes up a huge part of the variance of the susceptible class as we'd expect, with a smaller value of μ increasing our variance[8] .

For s, if we keep γ and μ constant, we see that as β rises, causing R_0 to rise, our variance increases, indicating that higher β causes values for s to deviate further from s^*. For i^* we don't see a relationship that's as clear. However, it turns out the effect of β is minimal when looking at our parameters. If we fix our values for γ and μ, we can look at $\frac{\partial \text{Var}(i^*)}{\partial R_0} = 1 - \frac{\mu}{\frac{\gamma}{7}+\mu}\left(\frac{1}{R_0^2} + \frac{2}{R_0^3}\right)$. For our purposes, we need only consider $R_0 \geq 1$. For all values of $R_0 \geq 0$ we see $\frac{\partial \text{Var}(i^*)}{\partial R_0} \approx 1$, indicating variance rises almost linearly as β increases[9].

4.2.3 Estimating time spent extinct with the Gaussian error function

One idea we can look at is use of the Gaussian distribution to approximate the amount of time spent extinct. For this approximation, we first consider plotting the level of infection for each time step in our simulation, similar to our SI plots such as 5. We now focus on the amount of infectious individuals for each point. We assume that the amount of infected individuals at any given time is normally distributed, with i^* being our mean and $c = \left(\sigma^2\right)_{2,2}$ our variance. From this assumption we can then create an estimate of what proportion of these normally distributed infected points would end up below $i = 0$. Obviously in a typical SIR style model there is never any points where $i = 0$, so we take all these points as points where I would be extinct. We therefore only need to know the equilibrium proportion of infected individuals i^*, as well as the variance c of these individuals (which we have just calculated using the covariance matrix) to find the

[8] A small μ corresponds to a longer life expectancy

[9] $\frac{\partial \text{Var}(i^*)}{\partial R_0}$ is so close that at $R_0 = 1$, $\frac{\partial \text{Var}(i^*)}{\partial R_0} = 0.999$ - and for larger R_0 it only gets closer

proportion of time the disease spends extinct. As we know that the points are Gaussian distributed. Therefore we have the following probability density function.

$$f(i) = \frac{1}{\sqrt{2c\pi}} \exp\left(-\frac{1}{2}\left(\frac{(i-i^*)}{\sqrt{c}}\right)^2\right)$$

We want to find Prob $(i \leq 0)$, so we must integrate over our probability density function, giving:

$$\text{Prob}(i \leq 0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2c\pi}} \exp\left(-\frac{1}{2}\left(\frac{(i-i^*)}{\sqrt{c}}\right)^2\right) di$$

We can think of this in terms of the Gaussian error function (erf), where:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x (e^{t^2}) dt$$

We can therefore rewrite Prob$(i \leq 0)$ in terms of the Gaussian error function to get the following formula.

$$\text{Prob}(i \leq 0) = \frac{1}{2}\left(1 - \text{erf}(\frac{i}{\sqrt{2c}})\right)$$

The Gaussian error function has many great approximations, so through using Python we can very easily find numerical values for this expression, giving us a sound approximation for the amount of time a disease spends extinct. We can use the following values we took from sections 4.1 and 4.2.2 to find this approximation.

$$\begin{aligned} i &= \tfrac{I^*}{N} = \tfrac{7\mu}{\gamma+7\mu}\left(1 - \tfrac{1}{R_0}\right) \\ c &= \tfrac{R_0-1}{N} + \tfrac{\mu(1+R_0)}{N{R_0}^2\left(\frac{\gamma}{7}+\mu\right)} \end{aligned}$$

From this we can predict the amount of time spent extinct as $N \to \infty$. This is because the Gaussian error function is an increasing function and as $n \to \infty$, erf$(n) \to 1$ and as $n \to -\infty$, erf$(n) \to -1$. Note that i does not change with different values of N, whereas c contains a factor of $\frac{1}{N}$. Therefore as $N \to \infty$, $\frac{-i}{\sqrt{2c}} \to -\infty$, so Prob$(i \leq 0) \to 0$. This kind of behaviour is exactly what we would hope would happen in our models. As our population sizes increase, we see the relative variance decrease, causing extinction events to become more rare.

5 The results

5.1 Looking at resonance in our model

5.1.1 The Fast Fourier Transform

The first thing to look at is resonance. For the implementation of this into my code I took heavy influence from the approach presented in [34] [10]. The most obvious way to look at any potential resonance in our model is to perform a Fast Fourier Transform. This is what I've done on the last 40 years of a 50 year simulation. Figure 6 shows a Fast Fourier Transform performed on some infection data generated from our Tau Leap model. First thing to note is that as our model operates with $\tau = \frac{1}{7}$. As one can see there are some larger peaks towards the smaller periods and as you get larger the power of any given peak starts to taper out, which is what one would expect. In real world systems it is common to see smaller frequencies of 7 days present [34] due to the weekly behaviour patterns in humans and how data is collected. These kinds of biases aren't present in our model so they're not to be expected. However some kind of frequencies around other parameters, such as average infectious period being visible on our plot, would make more sense.

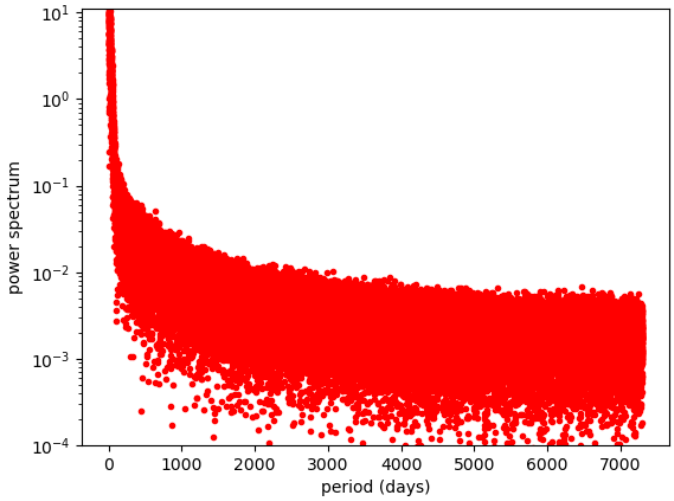


Figure 6: Output from our Fast Fourier Transforms for 10 runs averaged, with the x axis showing the day cycle and the y axis the power spectrum. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \gamma = 1, \tau = 1, N_0 = 2^{18}, R_0 = 5$

[10] Note as of the time of writing, the Github page links on this paper do not work anymore. One will now have to find them under Yoshiyasu Takefuji's new Github 'y-takefuji'

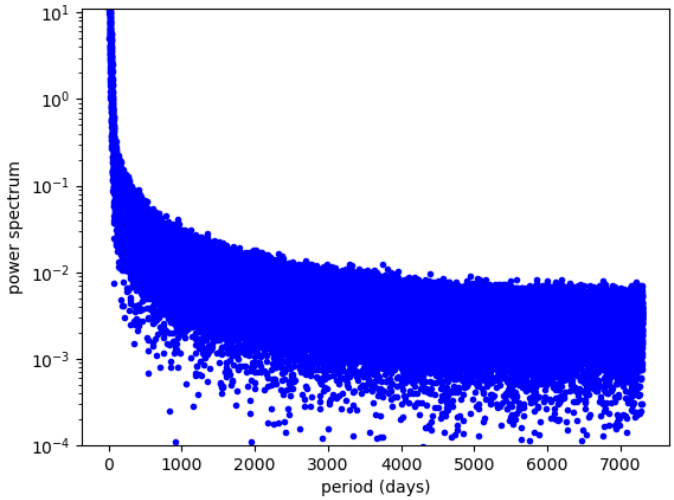


Figure 7: Output from our Fast Fourier Transforms for 10 runs averaged, with the x axis showing the day cycle and the y axis the power spectrum. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \gamma = 1, \tau = 1, N_0 = 2^{18}, R_0 = 5$

The two plots are very similar. However, it is worth noting that the points for the susceptible class for a given period tend to have less variability in their power, with us seeing a tighter band of points.

We see that both plots are very much dominated by the lower periods, whereas the higher periods each seem to have very little effect. This could be seen to suggest that the model either has very little periodicity (something which seems not to be true given Figure 5), or that our periodicity is mostly over a short period and therefore likely driven by high rate events such as infection or birth. However, although these high rate events have a great impact on the model, we know that the periodicity show in Figure 5 is driven by much lower frequencies. It is likely our Fast Fourier Transform approach has lost some of this detail, only showing us the very high power high frequency events.

5.1.2 Wavelets

When one looks at Figure 4, the periodicity one seems to see is over a much greater time period than days. As a result we're also interested in the averages over these time periods. However we're failing to see these kinds of frequencies appear. Therefore, it appears that I must use a tool more subtle than the Fast Fourier Transform. Inspired by an approach used by Grenfell[20], I have decided to opt for using wavelets.

For a Fast Fourier Transform, we look at the periodicity over the whole time period we're studying. Wavelets, however, can allow us to evaluate what frequencies we have present at every single timestep. This is particularly useful for systems where the natural

frequencies may change over time. Obviously for our system our rates are greatly affected by the numbers of individuals in different classes, as well as our total population, so the use of wavelets makes more sense than just applying a Fast Fourier Transform. When using wavelets, rather than utilising sinusoids as with the Fast Fourier Transform, we must decide what wavelet function to use. One option used in [20] was the Morlet wave function. I decided to test around 20 different wavelets to see which one did the best job of picking up our signals. Although the Morlet wavelets were able to pick out all our frequencies, I found that the local wavelet power spectrums produced by complex wavelets appeared far clearer than those produced by non-complex wavelets. In Figure 8 one can see the local power spectrum produced by a non-complex first order Gaussian derivative wavelet. Finally, I found that the wavelet representing the first order derivative of the complex Gaussian function $C \exp^{-it} \exp^{-t^2}$, where C is a constant, produced both the clearest local wavelet power spectrum, along with the smoothest global power spectrum.

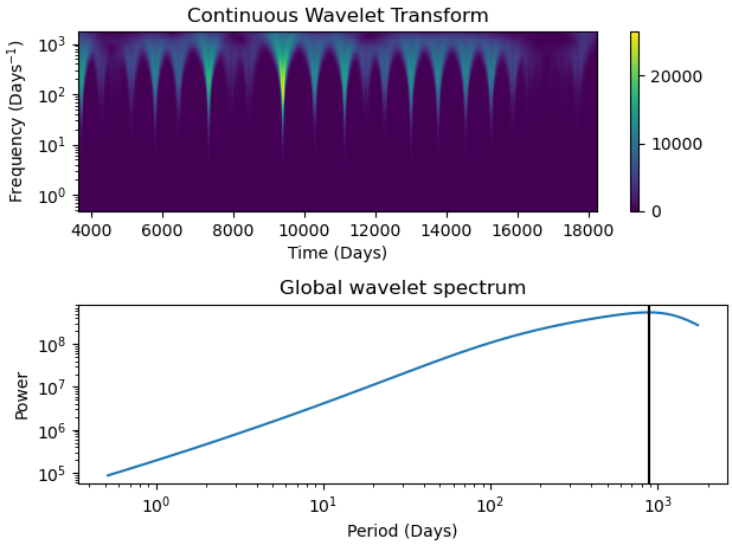


Figure 8: Output from using Wavelets with the first order derivative complex Gaussian wavelet as described. The black line at around 900 represents the dominant period. $\mu = \frac{1}{50*365}, \gamma = 1, \tau = 1, N_0 = 2^{18}, R_0 = 5$

In the global power spectrum we observe clear spikes as time goes along. These spikes correspond with the spikes in infection we see in graphs such as Figure 4. The larger spikes in infection produce larger power frequencies - illustrated by the yellow tones on our local wavelet power spectrum in Figure 8.

The global power spectrum is somewhat analogous to our Fast Fourier Transform, except it is far better at picking up our larger signals that may usually be cancelled. As one can see, we get a smooth global power spectrum compared to what we may find using real world case data [20], with a clear rise and peak. This peak appears at around the 900

in this case and is illustrated by the black line. The peak is dependent on imports and is often around 10^4. However the location of the peak varies depending on our level of imports, with lower imports causing the peak to have a higher period . One may wonder what this peak corresponds to, and a logical assumption would be that it corresponds to imports in some way. We can compare it to the average length of each wave of the disease following reintroduction. We find in this case it's 1100 and in the same order of magnitude. However, our results also match up with what was calculated in 4.2.1, where we found a period of 1200 for the same parameters. This was based on a simplified model that did not include imports. This suggests that, although the periodicity is continued by our imports (otherwise the disease would go extinction forever), the exact period we get is due to our choice of parameters for the system.

5.2 Effect of changing population size and the amount of exposure/infection classes

My first choice was to test the relationship between population size with extinction rate and length, as well as the differences we see when changing the number of exposure/infection classes. Throughout this I kept the average period individuals are infected at a constant (1 week) and simply changed the number of infection classes. For each number of classes I then varied the number of individuals in the population.

Next, we will look at plots 9 and 10 comparing the number of individuals in a population with disease extinction rate and average extinction length for each of the 3 different types of disease we're testing. For each of these we include 98% confidence intervals, which are indicated by the shaded areas. Each of these plots is based on 50 runs of the disease for 50 years with the first 10 years removed to account for the transitory stage we often see in stochastic models.

For the models themselves I decided to set

$$\text{average infectious period} = \frac{\text{Number of infected states} - \text{Number of exposure states}}{\gamma}$$

Then $\beta = \frac{R_0}{\text{Average infectious period}}$. This formulation ensures that our value of β has little effect on the model as R_0 can be kept the same for all 3 diseases.

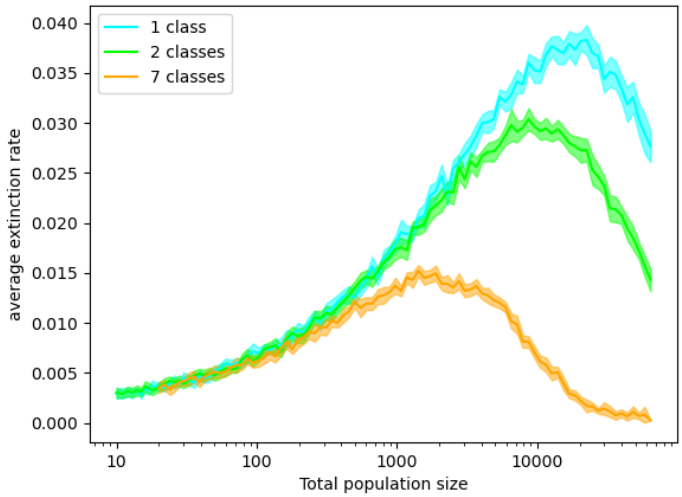


Figure 9: average extinction rate for each number of infection classes at varying population sizes, with 98% confidence interval error bars included. Note the x-axis has a log 10 scale. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = 1, R_0 = 5$

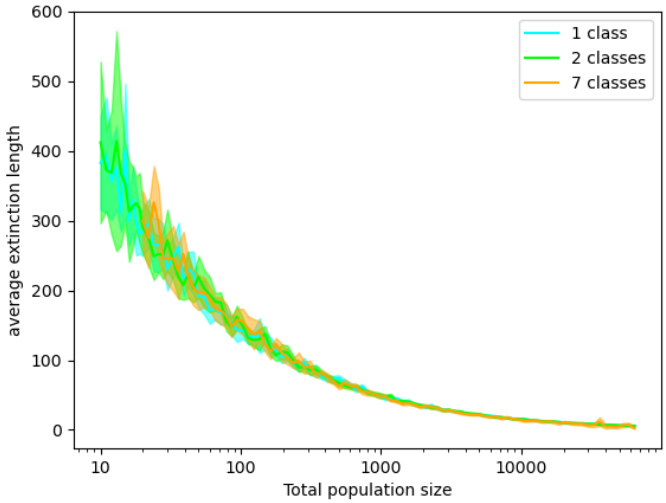


Figure 10: average extinction length for each number of infection classes at varying population sizes, with 98% confidence interval error bars included. Note the x-axis has a log 10 scale. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = 1, R_0 = 5$

From each model we have a somewhat surprising outcome. For 9 we see some clear difference between the 3 different diseases. We see for very low populations the relation between extinction rate and population size is much the same for the disease regardless of the number of infected classes. This may seem counter-intuitive at first, as a larger population surely makes the disease harder to eliminate. However one must remember

that the extinction rate is $\frac{\text{Number of extinctions}}{\text{Time}}$. For us to have a high extinction rate, the disease must become extinct and then be reintroduced shortly afterwards for the cycle to repeat. If the disease gets rarely reintroduced not many extinctions can even take place, so our extinction rate will have to be low. But if one remembers the formulation of the system from section 3.1, we have imports that scale with $\sqrt{N}$, hence as population size grows our extinction rate increases - at least at first.

For each of the 3 we see a clear peak in extinction rate corresponding with a particular population size. We also see for a smaller number of infection states, it tends to peak at a larger population and has a much higher peak for the rate of extinction. As our number of infection classes increases, not only does our peak extinction rate decrease, but we also find that it's achieved at a much lower population.

After these peaks, the effects of having a larger population becomes more pronounced than the effect of imports. The larger population is now starting to make it difficult for the disease to reach 0 infected individuals.

However, the effect of the amount of disease classes on the average length of extinction seems to be minimal, as shown on Figure 10. For each of the 3 diseases, we see near identical curves showing a decrease in average disease length as the total population size rises. This makes intuitive sense, as for larger populations we have a higher amount of imports, which will mean that any extinctions will likely happen for a shorter length of time.

One may wonder if these effects are simply due to the odd choice of class sizes for the model. Although I kept R_0, and the average time spend in all our infectious and exposure classes constant, it is possible that something such as the ratio of infectious and exposure classes is driving this behaviour and not the actual classes themselves. However, I repeated this for 3 diseases with 2, 4 and 8 infected classes respectively. For each of these diseases, exactly half of each classes are exposure classes vs infectious classes. The results show the exact same trend - showing that this isn't the case. Plots of these can be seen in the appendix with Figures 20 & 21.

The difference in the results for each 3 models is due to the difference that adding these extra classes does to the dynamics of the system. In section 3.1 I mentioned that when we split our model into multiple classes, we affect how the time spent in the infectious period is distributed. This is likely the cause of the differences we see in extinction rate for each population. What is interesting, however, is that this has no effect on the average extinction lengths.

5.3 Effect of changing average time infected

The next question to ask, is given a fixed number of infectious classes, what is the effect on changing the average length one stays in the infectious period. Below are Figures plotting extinction rate and average extinction length against average infectious period, with R0 = 5, 10 & 20 present. For each point on our graph we have $\beta = \frac{R_0}{\text{average infectious period}}$ - the same formulation as what we did for infectious classes. μ is kept constant for each of these plots, the only thing that changes is our values for γ.

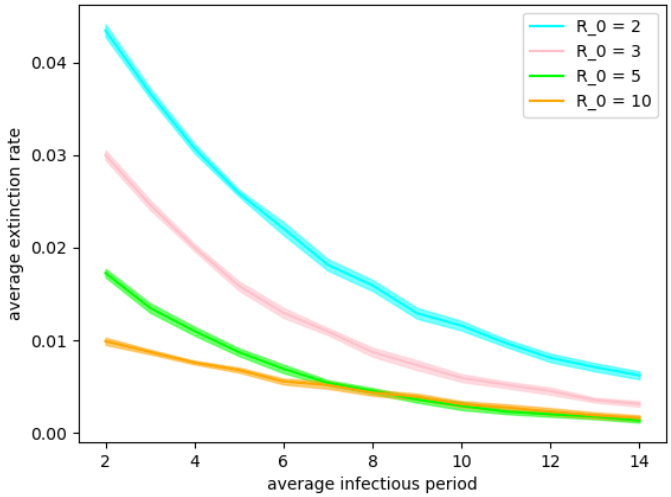


Figure 11: Average extinction rate for varying infectious periods at varying R_0, with 98% confidence interval error bars included.Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, N_0 = 2^{14}$

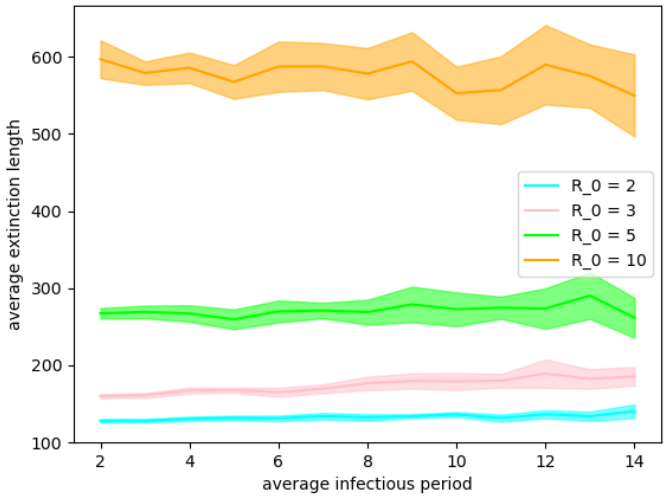


Figure 12: Average extinction length for varying infectious periods at varying R_0, with 98% confidence interval error bars included. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, N_0 = 2^{14}$

We can see from Figure 11 that for all R_0 values that there is a strong negative correlation between average infected period and average extinction rate. This is intuitive, as if we consider a scenario where there is only 1 infected individual in the whole population, a longer averaged infected period means that before the individual recovers or dies we have a greater chance of an import as more time will have passed as compared to a shorter period. Note the chance of an additional infection occurring due to our individual over the course of their infection is actually the same for both average infection periods as R_0 is fixed, so in this case imports are the main affecting factor.

We can also see that it appears as if for each of our R_0 values, when infectious period $\to \infty$, extinction rate $\to 0$ which makes intuitive sense with what has been stated above.

Next we look at Figure 12. For this we can see that the average infectious period seems to have no impact on average extinction length. Instead our main driver is our values of R_0. I will go into more detail on the relationship in the section 5.4

The extinction lengths are near constant, as the extinction rates decrease with average infectious period, indicating our time spent extinct for each value of R_0 decreases as the infectious period increases. We can look to our Gaussian error method to see if that agrees with this trend. We find that as the average infectious period rises, γ gets smaller. If we

look at our values for i^* we see that:

$$\begin{aligned} i^* &= \frac{7\mu}{\gamma+7\mu}\left(1-\frac{1}{R_0}\right) \\ &= \frac{7\mu}{\gamma+7\mu}-\frac{7\mu}{3\beta} \end{aligned}$$

So as γ gets smaller, our values for i^* get larger. For variance we see that:

$$\begin{aligned} c^* &= \frac{R_0-1}{N}+\frac{\mu(1+R_0)}{N{R_0}^2\left(\frac{\gamma}{7}+\mu\right)} \\ &= \frac{3\beta}{N(\gamma+7\mu)}-\frac{1}{N}+\frac{\mu\left(1+\frac{3\beta}{\gamma+7\mu}\right)}{N\left(\frac{3\beta}{\gamma+7\mu}\right)^2\left(\frac{\gamma}{7}+\mu\right)} \\ &= \frac{3\beta}{N(\gamma+7\mu)}-\frac{1}{N}+\frac{\mu\left(\frac{3\beta+\gamma+7\mu}{\gamma+7\mu}\right)}{N\left(\frac{(3\beta)^2}{\gamma+7\mu}\right)} \\ &= \frac{3\beta}{N(\gamma+7\mu)}-\frac{1}{N}+\frac{\mu(3\beta+\gamma+7\mu)}{N(3\beta)^2} \\ &= \frac{3\beta}{N(\gamma+7\mu)}-\frac{1}{N}+\frac{\mu(3\beta+\gamma+7\mu)}{N(3\beta)^2} \end{aligned}$$

For our parameter values μ is very small, making $\frac{\mu(3\beta+\gamma+7\mu)}{N(3\beta)^2}$ have a far smaller impact on c^*, as opposed to $\frac{3\beta}{N(\gamma+7\mu)}$. Therefore

$$c^* \approx \frac{3\beta}{N(\gamma+7\mu)}-\frac{1}{N}$$

Therefore we see that variance rises as γ gets smaller. Both variance and the mean are approximately proportional to $\frac{1}{\gamma}$. This means that $\frac{i^*}{\sqrt{2c^*}}$ will increase as γ get smaller, indicating that we get a reduction in the amount of time spent extinct. This matches well with our findings from the simulations.

5.4 Effect of changing R_0

The next obvious question, is what is the effect of changing our values of R_0. For this I kept all factors equal, with $\gamma = 1$ and β redefined at every R_0 value using $\beta = \frac{R_0}{\text{Average Infectious Period}}$. As one can see the plots for varying R_0 and its effects of extinction rate and average extinction length in Figures 13 and 14 respectively.

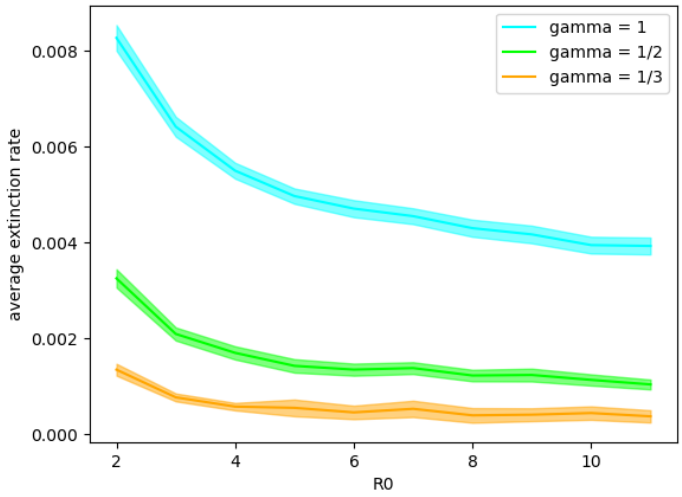


Figure 13: average extinction rate for different values of gamma at varying R_0, with 98% confidence interval error bars included. Note the we use a log scale for the y axis. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = 1, N_0 = 2^{14}$

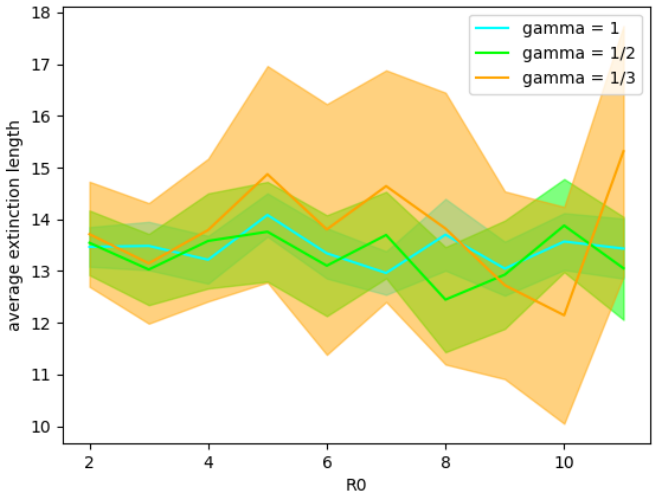


Figure 14: average extinction length for different values of gamma at varying R_0, with 98% confidence interval error bars included. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = 1, N_0 = 2^{14}$

As one can see for small values of R_0 in Figure 13, the effect of change in R_0 at small values results in a substantial decrease in extinction rate. However at around $R_0 = 4$ we see the effect of increasing R_0 ceases to be. One reason for this could be that for substantially small R_0, we may see a higher chance of a failure to invade and that after reaching a certain R_0 the change of failure to invade becomes so small that increasing R_0 does very

little. We then reach an equilibrium average extinction rate.

Meanwhile, we also see that the different values of γ, exhibit different equilbirium rates. This makes sense given our findings in 5.3. Regardless, they all exhibit very similar behaviour when reaching these rates.

In Figure 14 we see somewhat different behaviours. All 3 of our lines for each γ tend to go upwards with an increasing confidence interval as they go. This is intuitive, as a higher R_0 will result in an epidemic reaching its peak height far quicker and thus often results in a quicker extinction. However, the relation appears weak, as shown by the larger error bars.

Consequently, the time between the extinction and reinfections becomes larger and we tend to get a slightly longer average extinction length.

Interestingly, changes in γ seem to have very little effect here, as they all follow very similar trends. This suggests that the average infectious period has little effect on the average extinction length, a claim which is supported by Figure 12.

One final test to do is to use our estimate of the time spent extinct from the Gaussian error function as a point of comparison. The first thing to do is to see what happens to i^* and c^* when we change β. For i^* one sees that:

$$\begin{aligned} i^* &= \tfrac{7\mu}{\gamma+7\mu}\left(1-\tfrac{1}{R_0}\right) \\ &= \tfrac{7\mu}{\gamma+7\mu}\left(1-\tfrac{\gamma+7\mu}{3\beta}\right) \\ &= 7\mu\left(\tfrac{1}{\gamma+7\mu}-\tfrac{1}{3\beta}\right) \end{aligned}$$

So as β rises, i^* also rises proportional to $\frac{-1}{\beta}$

$$\begin{aligned} c &= \tfrac{R_0-1}{N}+\tfrac{\mu(1+R_0)}{NR_0{}^2\left(\frac{\gamma}{7}+\mu\right)} \\ &= \tfrac{3\beta}{N(\gamma+7\mu)}-\tfrac{1}{N}+\tfrac{\mu(1+R_0)}{NR_0{}^2\left(\frac{\gamma}{7}+\mu\right)} \\ &\approx \tfrac{3\beta}{N(\gamma+7\mu)}-\tfrac{1}{N} \end{aligned}$$

This indicates that as β rises, our variance also approximately rises proportionally. It will be challenging to understand exactly the behaviour of erf with both rising as R_0 does, so I will plot the behaviour. It is worth noting that at this size, the actual values we receive for the time spent extinct differ drastically from what we expect. This will be covered in more detail in section 6.2.1. Despite this, the trends we see as R_0 differs will remain the same.

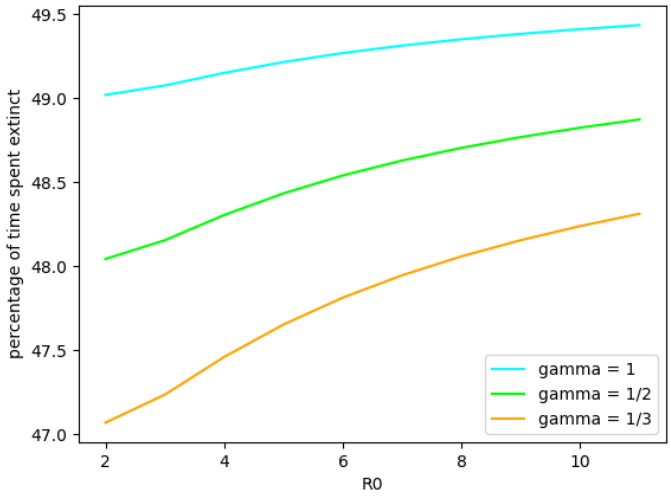


Figure 15: Calculated percentage of time spent extinct using our approximation based on the Gaussian distribution. $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, N_0 = 2^{14}$

However, this result doesn't match what one may imagine would happen to extinction time based off Figures 13 and 14. Figure 16 has therefore been plotted and shows the extinction times.

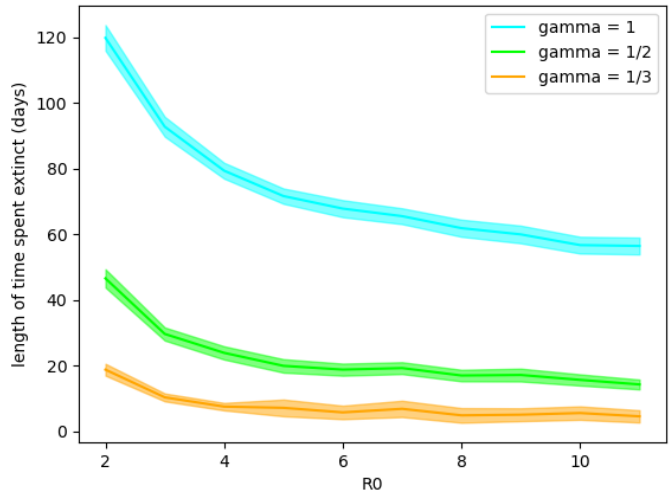


Figure 16: average extinction times for different values of gamma at varying R_0, with 98% confidence interval error bars included. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = 1, N_0 = 2^{14}$

As it turns out the use of the Gaussian based approximation has been a poor fit, as it contradicts the outcome from the stochastic model. The reason for this could be related to the limitations of such an approximation. Intuitively one might expect that as R_0 rises

we less failure to invade from imports, decreasing the amount of time the disease spends extinct. However one might also expect the extinction lengths to increase as R_0 rise, causing our disease to spend more time extinct. It appears that because the estimation of extinction time using the Gaussian distribution requires so many things to be simplified, it doesn't accurately reflect the effect changes in R_0 would have on the model. The weaknesses of the approximation will be discussed in more detail in section 6.2.1.

6 Discussion

6.1 The results

From our results we see some clear results for changes in population size on our extinction lengths and rates in Figures 9, 10, 20 and 21, with our extinction rate showing an increase, a peak and, then a decline, while average extinction lengths decline overall.

Bolker and Grenfell [4] found that as population grows, the 'fade out proportion' (the proportion of months with no cases) decreases. The reason they suggest for this is because a higher amount of infectives (due to a larger population size) decrease the effect of any stochasticity. We would expect extinction rate to follow a fairly similar pattern to this, and we see for large enough population sizes we do get a lower extinction rate, combined with a lower average extinction length, which would result in a decreased fade out proportion. The behaviour before the peak is a result of how imports are handled in this model. Further study could be conducted on the effect of changing the number of imports and seeing what happens to this peak, or perhaps even changing the way the model handles imports and seeing if the same peak still appears.

Average infectious period seems to also have a very clear effect, with a longer period decreasing the extinction rate, while having no effect on extinction length, resulting in less time spent extinct for high infectious periods. Our Gaussian distribution approximation agreed with this result. We also see that higher values of R_0 decrease the extinction rate, whilst increasing the average extinction length. It's also worth noting that for each of the 4 values of R_0, the extinction rates more close together as the infectious period increases. Given how clear the results have been for looking at average infectious periods, it would be interesting to see the effect on other common persistence parameters such as time to extinction.

There is an increase in extinction length as R_0 rises in Figure 14, as was suggested by Figure 12. Looking at extinction rate, as shown in Figure 13, we see a clear decrease as R_0 rises, which corresponds with what was suggested by Figure 11. However, when using the Gaussian approximation to calculate the time spent extinct we do not see this trend.

6.2 Future work

6.2.1 Weaknesses of the Gaussian distribution approximation

Although looking at our Gaussian distribution approximation for the amount of time spent extinct has been useful, it ultimately would have been beneficial to use it to get exact values for the percentage of time spent extinct at different parameter values. We can see that at our population levels, the values produced for the time spent extinct were inaccurate, as can be seen below

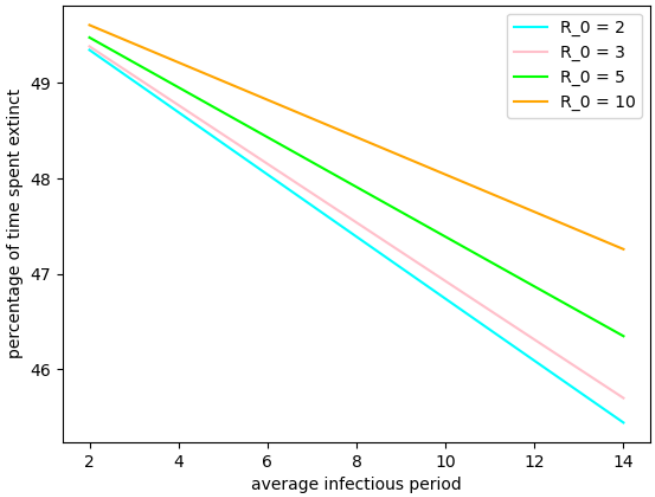


Figure 17: Calculated percentage of time spent extinct using our approximation based on the Gaussian distribution. $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, N_0 = 2^{14}$

Having a proportionally high variance as compared to our mean, is causing these incredibly high extinction rates, making the approximation far less accurate. If we run an instance of our model with the parameters $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, N_0 = 2^{14}$ and $\gamma = 1$, we get a mean that's similar to the value for our theoretical equilibrium point i^*. If we calculate i^* for these parameters we get 3.07×10^{-4}, whereas the actual mean for a simulation with these parameters ended up being 2.75×10^{-4}. However for our variance calculated from the results in 4.2.2 I got the variance as 2.44×10^{-4}, which is a relatively high variance given our mean. When running the simulation I found a variance of only 2.58×10^{-7}, which is significantly smaller.

There are several options for why the variances could differ so wildly from what we expect. The most obvious is a mistake in the calculation of the covariance matrix. However this has been checked thoroughly several times and is unlikely. The second option is that the covariance matrix is derived from an approximation of our model. We know for the stochastic versions, the SIR and SE^4I^3R models have their average time for an individ-

ual to spend in the infected class or classes follow an exponential or Erlang distribution respectively, and it may be that the variance produced from our computer simulation is smaller as a result. We also know that the approximation using the Gaussian distribution works best for large N, so it's possible that we aren't using large enough values to be able to make an accurate approximation.

Another reason why our approximation may be poor is because of use of imports. Obviously our model requires imports to ensure there's no permanent extinctions. However, if there are too many, the dynamics of our system move further away from the SIR simplification being used. If imports are too high, the disease will spend less time extinct, as an import means we have at least 1 individual in the system until they recover. Similarly we could have too few imports and the disease will spend far too long being extinct, as opposed to what our Gaussian based estimate suggests. Import levels will definitely have an impact on the amount of time the disease spends extinct. However, the Gaussian estimate does not reflect this in any way. It's unclear what the ideal number of imports are in order to make our distribution of infections in our simulation most reflect our Gaussian approximation. Further investigation on the impact of imports on the extinction process would be a logical next step.

The main reason as to why the Gaussian distribution approximation doesn't work is that it is strictly based on behaviour close to the fixed point i^*. As one moves further from the fixed point the nonlinear terms have a greater effect. For example, as we move away from i^* towards 0, we'd typically expect a higher number of susceptible individuals, making it easier for the disease to start infecting more individuals. This effect will only get worse as the disease moves further from i^*. As a result, the approximation simply does not match our results. We can even see how the distribution of infected individuals isn't Gaussian from Figure 18.

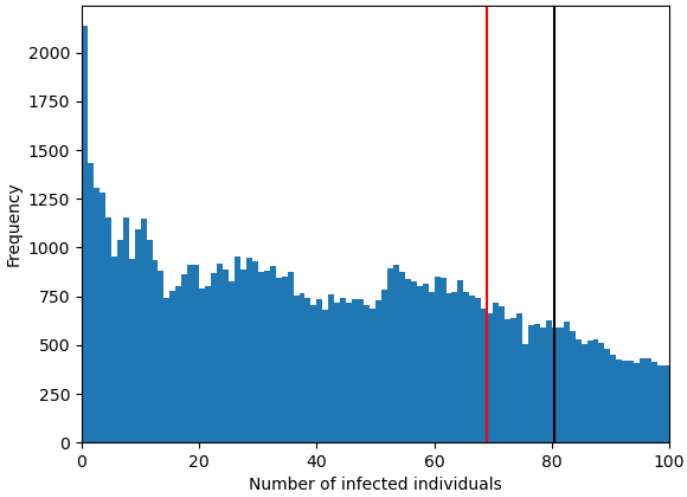


Figure 18: Frequency plot for the number of infected individuals in our system. The black line shows where i^* is located, whereas the red line shows the mean calculated from the data produced from this simulation. $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, N_0 = 2^{18}, R_0 = 5$

6.2.2 Variations on the model

As for much mathematical modelling, we have had to make several simplifications over the course of this dissertation. Many of these trace back to our deterministic equations, which may not reflect reality for most infections. Our deterministic equations assume perfect mixing of the population, which is rarely the case. Geographical location plays a huge impact as to who can get infected by whom. We must also note that for many diseases, once people are exposed, to a disease they may become symptomatic, causing their interactions with others to become limited. Import rates may also not be static, with certain times of year potentially producing more travel [11]. All of these factors have huge effects on the disease and are reasons why our deterministic equations could not reflect most diseases. However, for the case of studying disease persistence, I have chosen to ignore them as proper implementation with patch systems for geography, variable import rates, and having β vary between infection states would have greatly increased complexity, while not necessarily giving us to much further insight into what parameters affect extinction rate. However, looking at how the population is spread over patches as well as the contact rates between the patches affect the disease persistence could be interesting future work.

Another aspect of the project that could be explored further is the impact of very large population sizes. For most of the graphs such as Figure 11 each series would take around 24 hours to generate data for with a population of 2^{14}. Exploring higher population sizes

[11] We often see busy and less busy travel periods, so using some kind of sinusoidal function may be more appropriate

would mean that we could study extinction of low level diseases for very large population sizes, which may give us different outcomes.

6.2.3 Programming the model

To look at these higher values one would either need to change to a more efficient programming language than Python, use more powerful hardware, or use an even more efficient algorithm. Python is a high level language meaning that we could gain efficiency by using a lower level language such as C++. This would then allow us to explore outcomes for higher population sizes with ease, but would make coding the models far more challenging, as C++ is commonly considered a more complicated language than Python.

We could have made several choices for more efficient algorithms. In our model there's a requirement to have a small value of τ, due to how quickly an infection occurs. If our value of τ is too large, this can significantly disrupt the dynamics of the process. For example, if we set $\tau = 1$ our analysis outlined in section 5.1 completely changes, as now performing a fast Fourier transform on our total infected individuals over time gives us Figure 19.

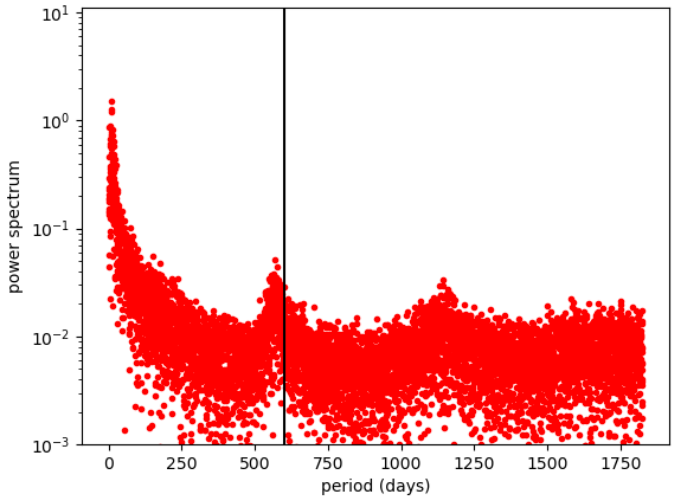


Figure 19: Output from our fast Fourier transform applied to average infections over 250 runs, with the x axis showing the day cycle and the y axis the power spectrum. In this case our value for τ is far too large. Only last 40 years out of 50 years of data used for each run. The black line represents a period of 600 days, where we seem to have a peak. $\mu = \frac{1}{50*365}, \tau = 1, N_0 = 2^{18}, R_0 = 5$

These dynamics with our multiple peaks completely disappear after we reduce our values of τ to $\frac{1}{7}$, which indicates that this behaviour seems to be caused by inappropriate τ value, which causes a large deviation from the dynamics presented in the deterministic equations. This complete change in dynamics may be due to how a high value for τ can

distort probabilities. For example, with the Tau Leap algorithm we assume that events are Poisson distributed. However, consider an event with rate 1 and $\tau = 1$. We would expect on average this event to take 1 time length, however as, $\tau = 1$, our event is bounded to take a minimum of 1 time length, but still may take more, which completely distorts the probabilities in our simulated models as compared to the deterministic equations. If one combines this effect with several population classes, individuals are meant to move quickly through (our exposure and infectious classes), and unsurprisingly, the behaviour of the model was greatly changed given initially $\gamma = 1$ and $\tau = 1$.

This error was caught late into the project, as the peaks seem to align with the large time frames we might expect from imports - so the idea of something being wrong hadn't occurred to me. However I was able to find similar by using wavelets. I am curious why the distribution being distorted in this way seemed to make the behaviour occur in the fast Fourier transform, as opposed to just the wavelets method - this could be an area of future exploration. I am unaware of any significant bodies of research on what happens to $\mathrm{SE}^m\mathrm{I}^n\mathrm{R}$ models as τ varies, and although this may contain limited direct real world benefit, further insight into τ-leaping may provide further insight into how the algorithm can be improved.

We must also note that although very small τ value means our model is closer to what we'd get from Gillespie (and thus better described our deterministic model), a very small τ makes our modelling take far longer as opposed to a larger τ value. In one run of our model with $N_0 \approx 50,000, \mu = \frac{1}{5*365}, R_0 = 5, \tau = \frac{1}{7}$ there were $2,525,571$ events, of which $262,892$ were births, $262,158$ were deaths and 201 were imports. Every other event in our system requires members of the infected classes to occur and as seen by Figure 4, these events do not happen frequently. Therefore it seems that we can make efficiency improvements in our model, particularly surrounding these "quiet" periods, where there are no infected class related events, while still capturing dynamics relatively close to the deterministic equations.

One way I could've done this was using the R-leaping algorithm, which is outlined in [2]. The R-leaping algorithm employs a similar approach to τ-leaping, but instead of using a set time length uses a set number of events S to simulate at once. This could prove far more efficient, with a clever choice of S allowing us to spend less computational resources on the uneventful infection free periods, while potentially getting more accurate results for the periods where we see infection. The only issue would be not picking S so small as to have no efficiency gains, but not so big such that the effect of any import only happens after a significant amount of time when we start simulating the next batch of events.

One final consideration I could've made in regards to choice of algorithm is picking an adaptive algorithm such as adaptive τ-leaping [5], adaptive R-leaping [2], or S-leaping [30]. We know that most of the time the system outlined in our original set of deterministic equations acts relatively stiffly, except for whenever we have any amount of individuals in the infected class. We could look at having variable values of τ depending on if we have the presence of any infected individuals, which could've been ideal for modelling extinctions.

6.2.4 Testing the model

In my testing phase to collect extinction rate and length data, I ran the τ-leap model multiple times [12], for 50 years each, discarded the first 10 years for each run and then collected data based from these runs. These runs could often take well over 70 hours[13], so in future testing may have to change. One question that immediately comes to mind is if we need the 10 year burn in. If I were to remake the project I would have tried to find some way to determine the burn in time required.

However, burn in may not even be required to begin with. For Monte Carlo models like the stochastic system of equations, the main reason burn in is used is to avoid using a potential transient stage as a part of our data. However if we simply pick a point the system has a high likelihood of being on, we also avoid the transient stage while avoiding a large amount of simulation. Such a point could potentially be calculated manually (likely using an approximation of our system of equations), or it could be found by simply counting how many times the system appeared at each point during the last simulation and then starting the next simulation on that exact point.

7 Conclusion

To conclude, we see that parameters such as the average infectious period and population size have a clear impact on the extinction rate. We also discovered that although it initially had a big affect after a certain point increasing R_0 has a very limited effect on our extinction rate.

One of the more surprising discoveries found was the effect the amount of exposed/infectious classes had on our extinction rate. We had almost no difference in the average extinction length, whereas the different division of classes affected the population rate curve greatly. This is likely due to how having more classes affects the underlying distribution of infected periods, thus shifting the dynamics of the disease.

[12] generally 100 times depending on how long it would take
[13] On a 4 core laptop released around 2019

Bibliography

References

[1] David Alonso, Alan J McKane, and Mercedes Pascual. "Stochastic amplification in epidemics". In: *Journal of the Royal Society Interface* 4.14 (2007), pp. 575–582.

[2] Anne Auger, Philippe Chatelain, and Petros Koumoutsakos. "R -leaping: Accelerating the stochastic simulation algorithm by reaction leaps". In: *The Journal of Chemical Physics* 125.8 (Aug. 2006). ISSN: 1089-7690. DOI: 10.1063/1.2218339. URL: http://dx.doi.org/10.1063/1.2218339.

[3] Scott Barrett. "Economic considerations for the eradication endgame". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (Aug. 2013).

[4] Benjamin Bolker and Bryan Thomas Grenfell. "Space, persistence and dynamics of measles epidemics". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 348.1325 (1995), pp. 309–320.

[5] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. "Adaptive explicit-implicit tau-leaping method with automatic tau selection". In: *The Journal of Chemical Physics* 126.22 (June 2007). ISSN: 1089-7690. DOI: 10.1063/1.2745299. URL: http://dx.doi.org/10.1063/1.2745299.

[6] Victoria Chebotaeva and Paula A Vasquez. "Erlang-distributed SEIR epidemic models with cross-diffusion". In: *Mathematics* 11.9 (2023), p. 2167.

[7] Victoria Chebotaeva and Paula A. Vasquez. "Erlang-Distributed SEIR Epidemic Models with Cross-Diffusion". In: *Mathematics* 11.9 (2023). ISSN: 2227-7390. DOI: 10.3390/math11092167. URL: https://www.mdpi.com/2227-7390/11/9/2167.

[8] Lewis Cole. *Gillespie algorithm*. Apr. 2020. URL: https://lewiscoleblog.com/gillespie-algorithm.

[9] Andrew JK Conlan et al. "Resolving the impact of waiting time distributions on the persistence of measles". In: *Journal of the Royal Society Interface* 7.45 (2010), pp. 623–640.

[10] Ian Cooper, Argha Mondal, and Chris G Antonopoulos. "A SIR model assumption for the spread of COVID-19 in different communities". In: *Chaos, Solitons & Fractals* 139 (2020), p. 110057.

[11] CE Dangerfield, Joshua V Ross, and Matthew James Keeling. "Integrating stochasticity and network structure into an epidemic model". In: *Journal of the Royal Society Interface* 6.38 (2009), pp. 761–774.

[12] Paul L. Delamater et al. "Complexity of the basic reproduction number (R0)". In: *Emerging Infectious Diseases* 25.1 (Jan. 2019), pp. 1–4.

[13] B. Finkenstädt, M. Keeling, and B. Grenfell. "Patterns of density dependence in measles dynamics". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265.1398 (May 1998), pp. 753–762.

[14] Christopher Fitzpatrick et al. “The cost-effectiveness of an eradication programme in the end game: Evidence from guinea worm disease”. In: *PLOS Neglected Tropical Diseases* 11.10 (Oct. 2017). DOI: 10.1371/journal.pntd.0005922.

[15] Charlie Geyer. *Burn-in is unnecessary*. URL: http://users.stat.umn.edu/~geyer/mcmc/burn.html#meyn.

[16] Daniel T Gillespie. “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. In: *Journal of Computational Physics* 22.4 (Dec. 1976), pp. 403–434. ISSN: 0021-9991. DOI: 10.1016/0021-9991(76)90041-3. URL: http://dx.doi.org/10.1016/0021-9991(76)90041-3.

[17] Daniel T. Gillespie. “Approximate accelerated stochastic simulation of chemically reacting systems”. In: *The Journal of Chemical Physics* 115.4 (July 2001), pp. 1716–1733. ISSN: 1089-7690. DOI: 10.1063/1.1378322. URL: http://dx.doi.org/10.1063/1.1378322.

[18] Peter W Glynn. “Diffusion approximations”. In: *Handbooks in Operations research and management Science* 2 (1990), pp. 145–198.

[19] Peter W. Glynn. “Chapter 4 Diffusion approximations”. In: *Stochastic Models*. Vol. 2. Handbooks in Operations Research and Management Science. Elsevier, 1990, p. 145. DOI: https://doi.org/10.1016/S0927-0507(05)80168-9. URL: https://www.sciencedirect.com/science/article/pii/S0927050705801689.

[20] B. T. Grenfell, O. N. Bjørnstad, and J. Kappey. “Travelling waves and spatial hierarchies in measles epidemics”. In: *Nature* 414.6865 (Dec. 2001), pp. 716–723. ISSN: 1476-4687. DOI: 10.1038/414716a. URL: http://dx.doi.org/10.1038/414716a.

[21] Jifeng Hu, Hye-Won Kang, and Hans G Othmer. “Stochastic analysis of reaction–diffusion processes”. In: *Bulletin of mathematical biology* 76 (2014), pp. 854–894.

[22] Matt J. Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Sept. 2011. ISBN: 9780691116174. DOI: 10.2307/j.ctvcm4gk0. URL: http://dx.doi.org/10.2307/j.ctvcm4gk0.

[23] Matthew James Keeling and Joshua V Ross. “On methods for studying stochastic disease dynamics”. In: *Journal of the Royal Society Interface* 5.19 (2008), pp. 171–181.

[24] Petra Klepac et al. “Six challenges in the eradication of infectious diseases”. In: *Epidemics* 10 (2015), pp. 97–101.

[25] Petra Klepac et al. “Six challenges in the eradication of infectious diseases”. In: *Epidemics* 10 (2015). Challenges in Modelling Infectious DIsease Dynamics, pp. 97–101. ISSN: 1755-4365. DOI: https://doi.org/10.1016/j.epidem.2014.12.001. URL: https://www.sciencedirect.com/science/article/pii/S175543651400070X.

[26] Thomas G Kurtz. “Limit theorems for sequences of jump Markov processes approximating ordinary differential processes”. In: *Journal of Applied Probability* 8.2 (1971), pp. 344–356.

[27] Thomas G Kurtz. “Solutions of ordinary differential equations as limits of pure jump Markov processes”. In: *Journal of applied Probability* 7.1 (1970), pp. 49–58.

[28] Rachel Kuske, Luis F Gordillo, and Priscilla Greenwood. “Sustained oscillations via coherence resonance in SIR”. In: *Journal of theoretical biology* 245.3 (2007), pp. 459–469.

[29] Jia Li. “Persistence in discrete age-structured population models”. In: *Bulletin of mathematical biology* 50 (1988), pp. 351–366.

[30] Jana Lipková et al. “S-leaping: an adaptive, accelerated stochastic simulation algorithm, bridging τ-leaping and R-leaping”. In: *Bulletin of mathematical biology* 81.8 (2019), pp. 3074–3096.

[31] Alun L Lloyd. “Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics”. In: *Theoretical population biology* 60.1 (2001), pp. 59–71.

[32] Ingemar Nåsell. “On the time to extinction in recurrent epidemics”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.2 (1999), pp. 309–330.

[33] Pejman Rohani and Aaron A King. “Never mind the length, feel the quality: the impact of long-term epidemiological data sets on theory, application and policy”. In: *Trends in ecology & evolution* 25.10 (2010), pp. 611–618.

[34] Yoshiyasu Takefuji. “Fourier analysis using the number of COVID-19 daily deaths in the US”. In: *Epidemiology & Infection* 149 (2021), e64.

A Further figures

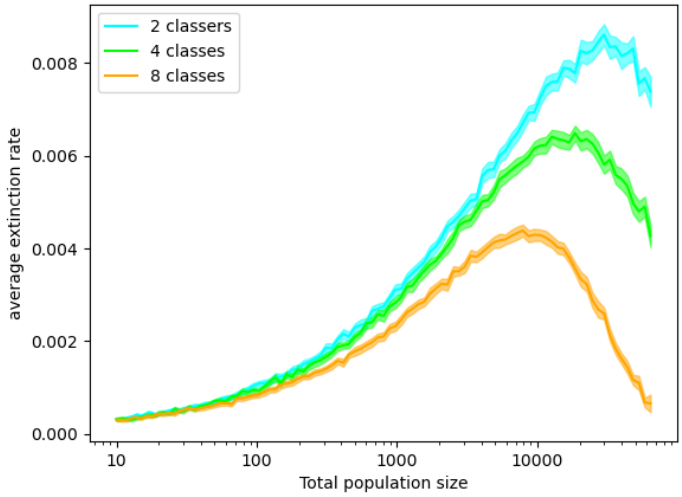


Figure 20: average extinction rate for each number of infection classes at varying population sizes, with 98% confidence interval error bars included. Note the x-axis has a log 10 scale. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, R_0 = 5$

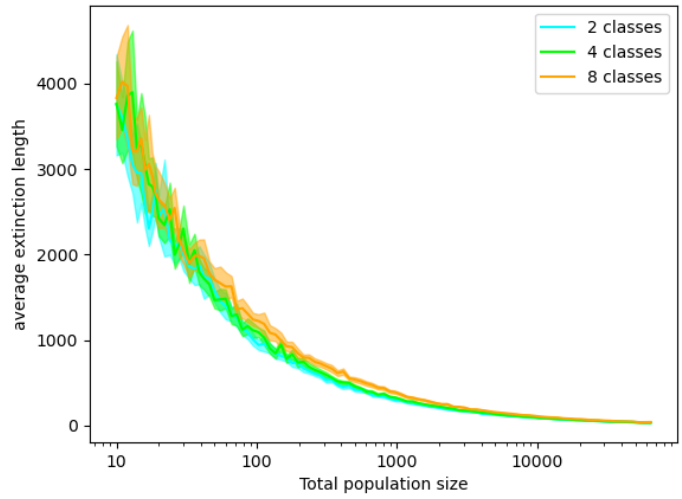


Figure 21: average extinction length for each number of infection classes at varying population sizes, with 98% confidence interval error bars included. Note the x-axis has a log 10 scale. Only last 40 years out of 50 years of data used for each run. $\mu = \frac{1}{50*365}, \tau = \frac{1}{7}, R_0 = 5$