

# Bachelor Dissertation

Kasper Engelen, Jonathan Meyer, Dawid Miroyan, and Igor Schittekat

University of Antwerp

**Abstract.** This document reports our findings regarding the final dissertation. The sections and subsections correspond to the assignments given to us. In this project we worked with a simulator called Stride, developed at the University of Antwerp. We explore various concepts within computational epidemiology through the use of this program.

**Keywords:** Computational Epidemiology · Dissertation

## 1 Introduction

Epidemiology is the study of how diseases behave in defined populations. Important aspects of the field include disease causation, transmission and outbreak investigation. While these subjects can be studied using real data, it is often desirable to be able to perform experiments in silico. This way, predictions can be made regarding potential outbreaks, the effects on the population, and effects of possible counter-measures. To that end, software is created that is capable of simulating a modelled disease in a modelled population.

The simulator used in this paper is called Stride, developed at the University of Antwerp. This paper will explore the capabilities of the simulator itself, like the role of stochasticity and performance under various conditions. It also researches the influence different parameters on the model like the  $R_0$  value, introducing a vaccination after a set period of time, and using different populations.

## 2 Simulation

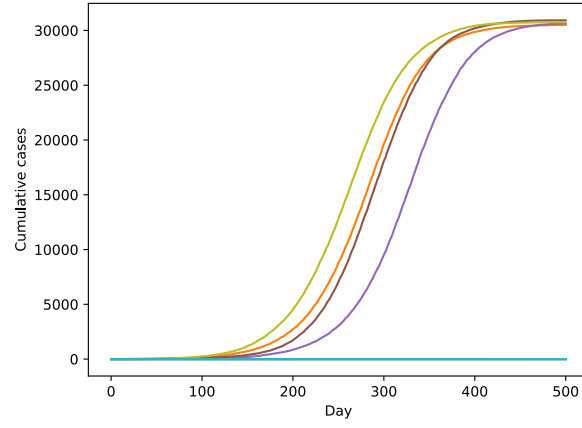
This section will primarily look at the effects of different conditions on the simulation results. First, we will look at the impact of stochastic variation on the results. Next, we will research outbreak behaviour by trying to classify scenarios as either 'outbreak' or 'extinction'. Finally, we will use the simulator to find estimates for the value of  $R_0$  and the population's immunity level using real data. This will show that Stride is indeed capable of simulating the modelled disease with decent accuracy, if we account for the (non-trivial) stochastic variation.

### 2.1 Stochastic Variation

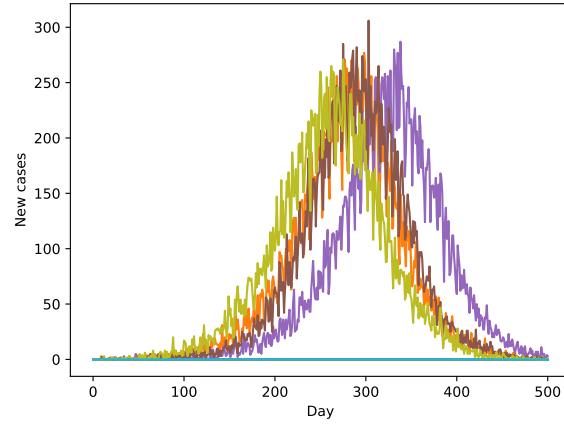
The first topic we consider is stochastic variation. Since the simulation uses a pseudo-random number generator, it's useful to inspect the influence of this

stochasticity on the results of the simulation. Using the Stan tool which is provided with Stride, we collected data on 100 simulations with an identical configuration file. The only difference between executions is the RNG seed. For the data we collected, we calculated a mean of 33.33 new cases per day, and a variance of 1196.99. Looking at the plot for the cumulative cases per day, two 'categories' can be distinguished: 'outbreak' scenarios and 'extinction' scenarios. These cases are quite evenly spread, which explains the high variance for new cases per day. In the former, the curve has a sigmoid shape, indicating that the disease successfully spread among the population. The latter scenario corresponds to the curves that are almost constant and are bounded by a value well under the population size. In these cases, the disease did not manage to spread, resulting in only a few infected people at the end of the simulation. This is likely explained by the initial infection: if the first person to be infected is sufficiently isolated (either socially or by being surrounded by people who are immune), the disease doesn't have a chance to spread. From the boxplot and a normal probability plot for the average new cases per day (Figure 2) (without extinction cases), it seems like the new cases per day do not have a normal distribution. However, the plot for new cases per day is very jagged (it jumps significantly from day to day), therefore this might not be representative.

Figure 1 shows the different scenarios. In the cumulative cases plot the 'extinction' scenario is clearly visible at the bottom. The plot for new cases per day is very jagged, which can also be explained by the stochastic nature of the simulation. From this we can conclude that stochastic variation has a significant impact on the simulation's outcome, and that a scenario should be simulated more than once to acquire representative results.

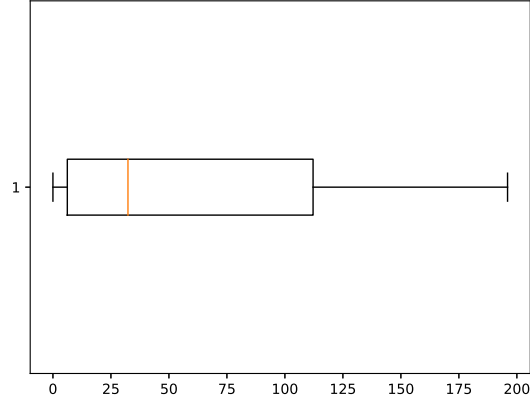


(a) Cumulative cases per day.

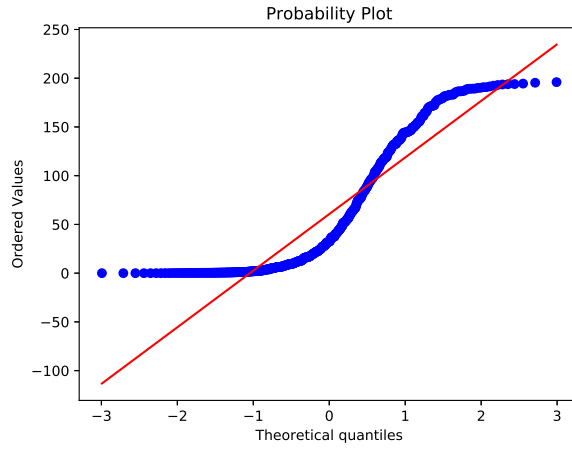


(b) New cases per day.

Fig. 1: Results of five different simulations, using a disease profile for measles. Seeding rate = 0.2%,  $R_0 = 11$ . Simulated for a population of 600,000, with 80% vaccine and immunity rate. This plot shows only five results for clarity, the trend is general and seen in any number of simulations.



(a) Boxplot for average new cases per day.



(b) Normal Distribution Plot for average new cases per day.

Fig. 2: Results of 100 different simulations, using a disease profile for measles. Seeding rate = 0.2%,  $R_0 = 11$ . Simulated for a population of 600,000, with 80% vaccine and immunity rate.

## 2.2 Extinction Threshold

As discussed previously, it might be the case that only very few people become infected over the course of a simulation. This is referred to as extinction. There is a clear distinction between outbreaks and extinctions, so in this subsection we attempt to find an *extinction threshold*.

Figure 3 gives the frequencies of the amount of infected people at the end of the simulations. The distinction between outbreaks and extinctions is quite clear from this histogram. There is one peak on the lower end of the x-axis, and there is a cluster on the higher-end. It is hard to find a good concrete value for the threshold, but the total amount of infected people in an extinction is always significantly under the total population size. Therefore, a fraction of the total population could be used as threshold. In our simulations the amount of cases after extinction never exceeded 50, so a threshold of 0.01% of the population seems reasonable.

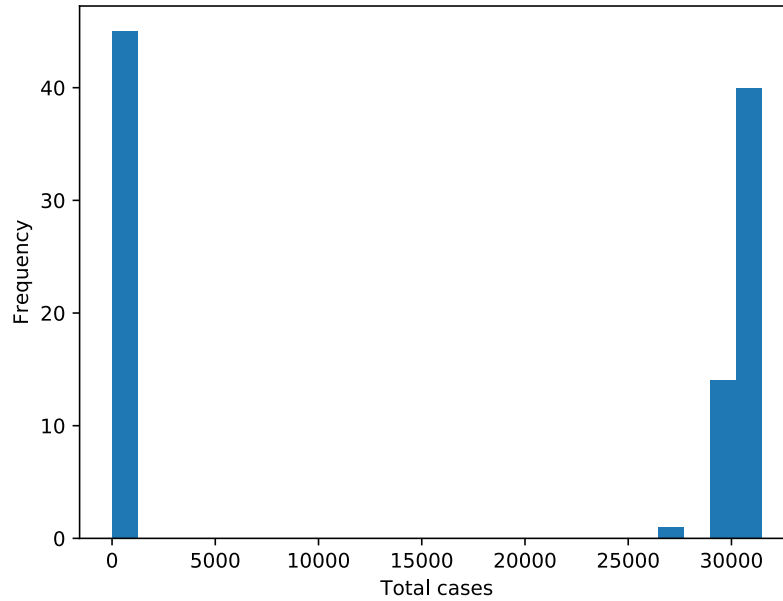


Fig. 3: Frequencies of final amounts of cases for 100 simulations, using a disease profile for measles. Seeding rate = 0.2%,  $R_0 = 11$ . Simulated for a population of 600,000, with 80% vaccine and immunity rate.

### 2.3 Immunity Level

In order to make assumptions about the population's immunity level, we look at the simulated results for different values. Upon experimentation, it becomes apparent the immunity level is approximately 70% of the population. Figure 4 shows the average new cases per day for different immunity levels. The reference

curve is also included. This plot was generated by taking the average of new cases per day over 20 simulations per immunity level. Using PyStride to simulate outbreaks, we narrowed down the immunity level  $I$  to  $0.705 \leq I \leq 0.7175$ .

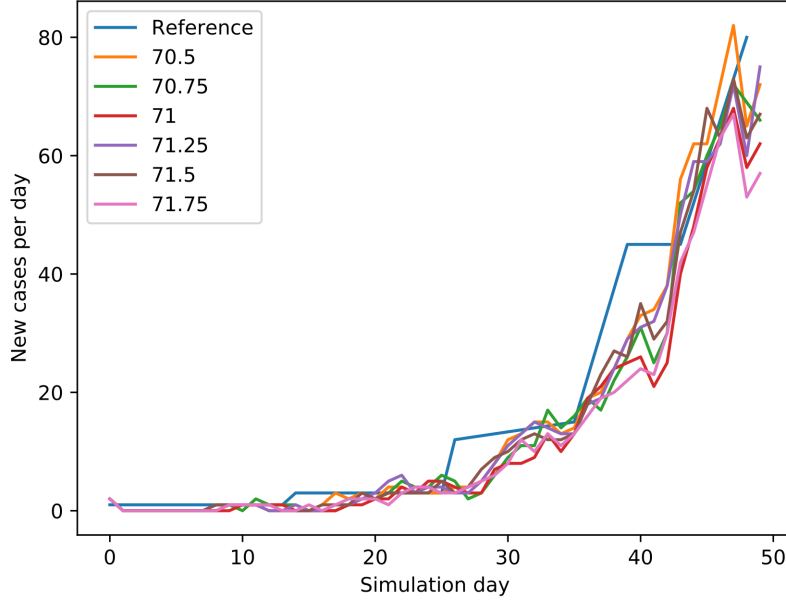
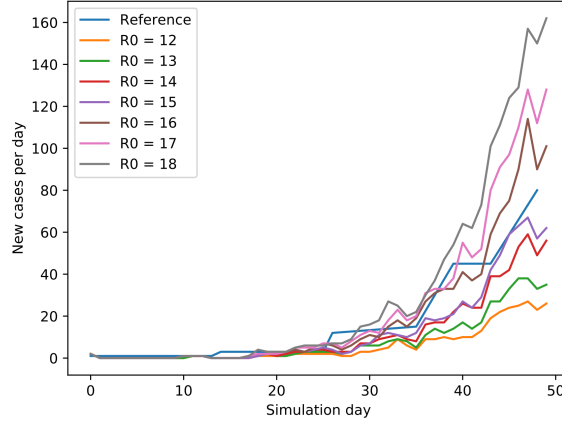


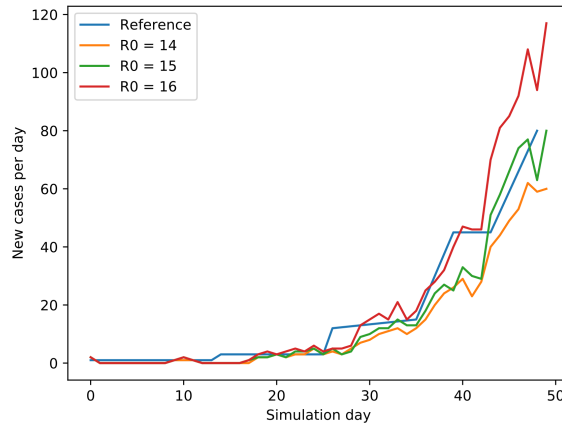
Fig. 4: Average new cases per day, calculated from 20 simulations per immunity level. Simulated using a disease profile for measles. Seeding rate = 0.000334%,  $R_0 = 15$ . Population size is 600,000, with random vaccine profile.

## 2.4 Estimating $R_0$

Now that we have a relatively decent estimate for the immunity level of the population, we can do an analogous exercise to estimate  $R_0$ . For this data, we fixed the immunity level to 70.5%. Figure 5 (a) shows the plot for each possible  $R_0$  in the range  $[12, 18]$  (average over 20 simulations). It's clear that the best candidates are 14, 15 or potentially 16. When we plot a more detailed plot for just those values (average over 30 different simulations), it seems the results for  $R_0 = 15$  are closest to the reference data, meaning it is likely an accurate value for the parameter for this modelled disease. As seen in Figure 5 (a), higher values ( $>16$ ) result in much more new cases per day (especially towards the end), and lower values ( $<14$ ) result in the opposite; a less drastic increase in new cases per day.



(a) Overview of all potential  $R_0$ 's. Averages from 20 runs per  $R_0$ .



(b) More detailed look at  $R_0$  candidates. Averages from 30 runs per  $R_0$ .

Fig. 5: Averages for different values of  $R_0$ .

### 3 Population generation

In this section we will investigate what impact the population structure and generation has on the simulations. More specifically we will investigate the impact of the age distribution, the effectiveness of the vaccination of college students, and whether or not commuting to work has an effect on disease spread.

### 3.1 Influence of demography

We want to determine whether or not the age distribution of a population has an impact on the size and probability of an outbreak. We start off with two populations, region A and region B, each generated based on a household file. The households mainly differ in age, with the households of region A being generally younger than those of region B. This is shown in figure 6.

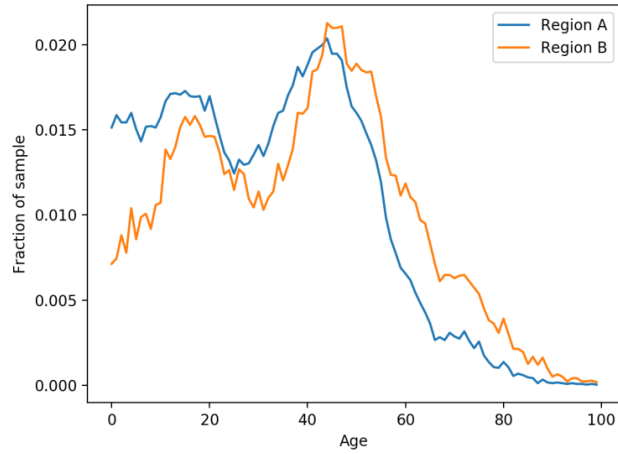


Fig. 6: The age distribution of region A and region B.

In order to see the impact of the difference in age distribution on the simulation, we had to run a number of simulations. Each simulation we would run had to be “interesting”, i.e. when compared to other simulations the impact of the age difference had to be clear. We listed a number of parameter sets so that all regions, big and small seeding rates, short and long simulation lengths, and the difference in weekdays would be covered.



Region	Seeding rate	Number of days	Weekday
Region A	0.002	50 Days	Monday
Region A	0.002	300 Days	Monday
Region A	0.0003	50 Days	Monday
Region A	0.0003	300 Days	Monday
Region A	0.00000167	50 Days	Monday
Region A	0.00000167	300 Days	Monday
Region B	0.002	50 Days	Monday
Region B	0.002	300 Days	Monday
Region B	0.0003	50 Days	Monday
Region B	0.0003	300 Days	Monday
Region B	0.00000167	50 Days	Monday
Region B	0.00000167	300 Days	Monday
Region A	0.002	50 Days	Saturday
Region A	0.002	300 Days	Saturday
Region A	0.0003	50 Days	Saturday
Region A	0.0003	300 Days	Saturday
Region A	0.00000167	50 Days	Saturday
Region A	0.00000167	300 Days	Saturday
Region B	0.002	50 Days	Saturday
Region B	0.002	300 Days	Saturday
Region B	0.0003	50 Days	Saturday
Region B	0.0003	300 Days	Saturday
Region B	0.00000167	50 Days	Saturday
Region B	0.00000167	300 Days	Saturday

Table 1: The different parameter sets that could potentially be interesting.

Next, we ran 400 simulations for each parameter set in table 1 and plotted different aspects of the simulations. These aspects include the outbreak sizes, the new cases per day, and the cumulative cases. We then used these plots to compare the two regions and visually selected plots that clearly showed the difference between the two regions. We then narrowed the sets in table 1 down to the 6 parameter sets listed below. Note that we also discarded the weekday column. This is because we didn't find any significant differences between Monday and Saturday.

Region	Seeding rate	Number of days
Region A	0.002	50
Region A	0.002	300
Region A	0.00000167	300
Region B	0.002	50
Region B	0.002	300
Region B	0.00000167	300

Table 2: The different parameter sets that we found to be interesting.

Some parameters used in the stride simulator are not listed in the table, since they were the same for all sets. All the simulations started on the 11th of March 2019, with  $R_0$  equal to 11 and a randomized vaccination policy with a rate of 0.8. *Seeding rate* here means the fraction of the population that is initially infected.

Now that we determined the six interesting parameter sets listed in table 2, we ran the final simulations on which we would base our conclusions. In order to make our calculations and conclusions statistically significant, we ran 1000 simulations per parameter set.

### Comparing outbreak sizes

Before analysing the outbreak sizes, we first filtered out all the simulations that did not lead to an outbreak. For the simulations with seeding rate equal to 0.002 the *extinction threshold* was taken to be 5000, for seeding rate equal to 0.00000167 we used 40000. We determined these thresholds by visually inspecting the plots. One should note that when setting the seeding rate to 0.002 there was a 100% outbreak rate and as such filtering out the extinctions was not really necessary.

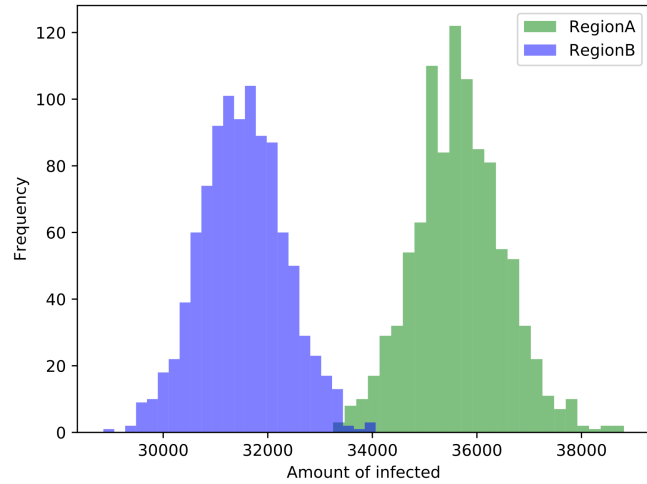


Fig. 7: Plot that compares region A and B with a simulation over 50 days with seeding rate 0.002.

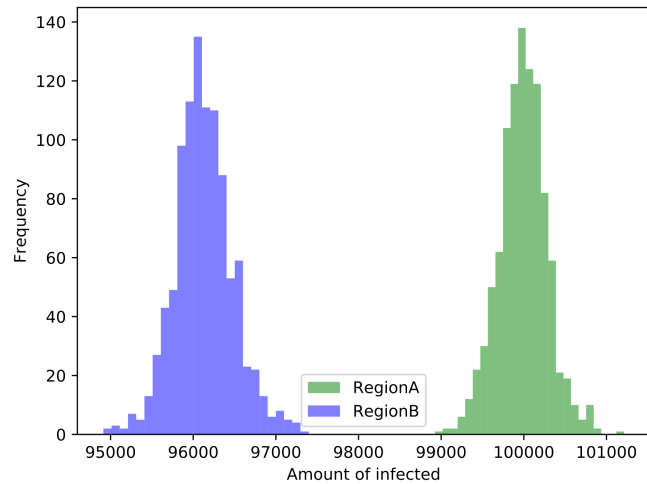


Fig. 8: Plot that compares region A and B with a simulation over 300 days with seeding rate 0.002.

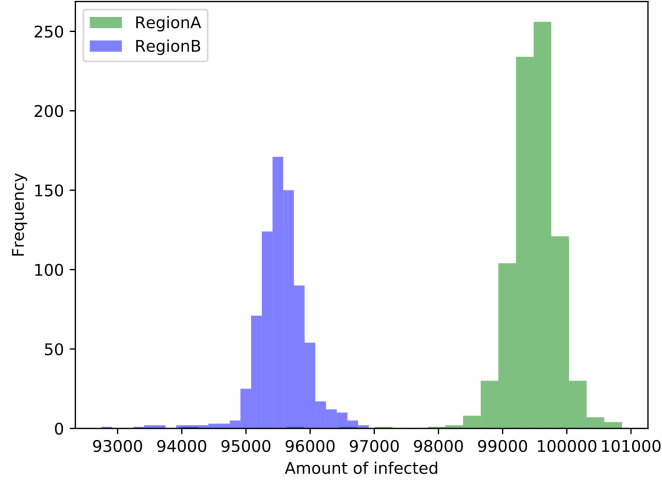


Fig. 9: Plot that compares region A and B with a simulation over 300 days with seeding rate 0.00000167.

First off, by visually inspecting figures 7, 8, and 9 we can see that the outbreaks in region A are consistently larger than those in region B. We will make a more formal statistical analysis based on the parameter set used in figure 9, due to the large outbreak size and visible difference between regions A and B.

We determined a 95% confidence interval for the outbreak size of region A:

$$[99449.975, 99510.976]$$

We did the same for region B:

$$[95198.107, 95521.279]$$

Since these intervals are disjoint and lie apart quite significantly, we can conclude that the outbreak size for region A is higher than that of region B.

### Comparing outbreak probabilities

In order to analyse the outbreak probability, we only looked at the simulations with seeding rate equal to 0.00000167, since with seeding rate equal to 0.002 there was a 100% outbreak rate. The following plot contains a histogram that depicts the amount of infected individuals in the different simulations.

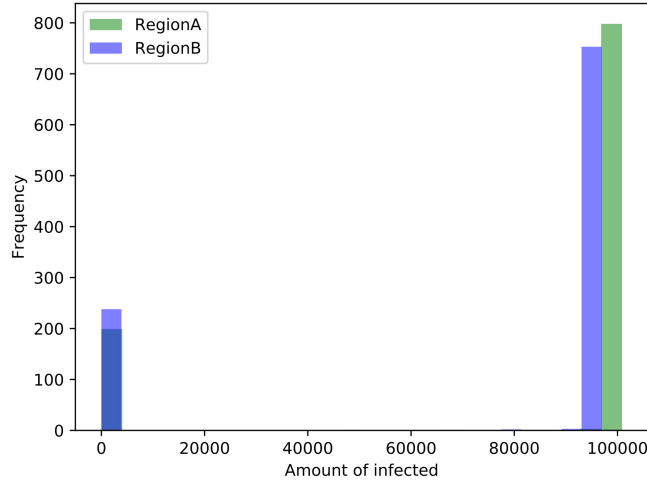


Fig. 10: Plot that compares regions A and B with a simulation over 300 days with seeding rate 0.00000167.

By visual inspection alone, we can see that there is a difference between region A and region B which leads to region A having a higher outbreak probability than region B. We also determined a 95% confidence interval for the outbreak probability of region A:

$$[77.625\%, 82.575\%]$$

As well as an interval of the same confidence for region B:

$$[73.561\%, 78.839\%]$$

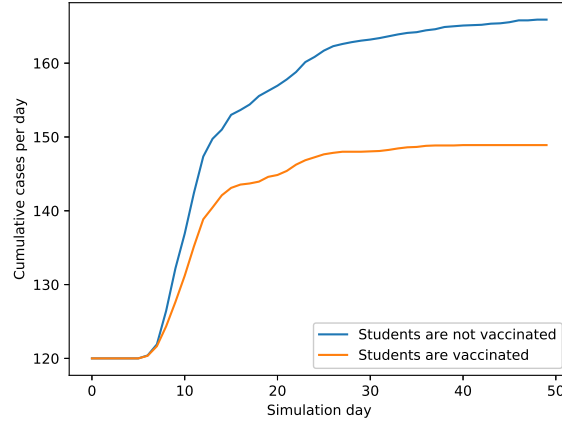
We can now conclude that the outbreak probability of region A lies higher than that of region B.

### 3.2 Vaccinating on campus

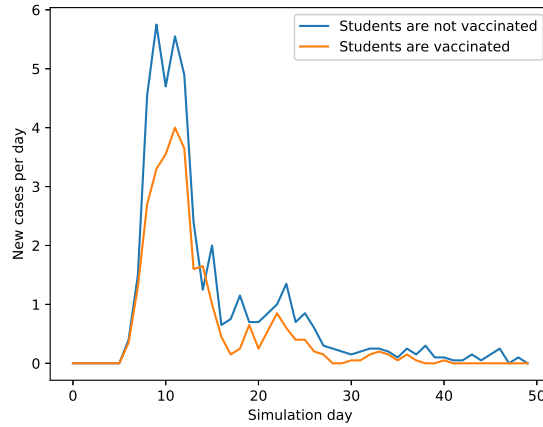
By vaccinating the students, quickly after infected individuals appear in the population, we want to test whether so called 'catch-up' campaigns have a noticeable impact on an outbreak.

For these simulations, we generated a population where 60% of people between the age of 18 and 26 attend high education. We then simulate scenarios where all students are either vaccinated on day 7 or not vaccinated at all. When looking at Figure 11, the amount of new cases is the same in both scenarios on day 5 and

6. As the students are not vaccinated in both scenarios, some of them become infected and can spread the disease. Starting from day 7, where in one of the scenarios the susceptible students are vaccinated, the amount of new cases is lower than when not vaccinating.



(a) Plot showing the amount of infected student per day.



(b) Plot showing the amount of newly infected student per day.

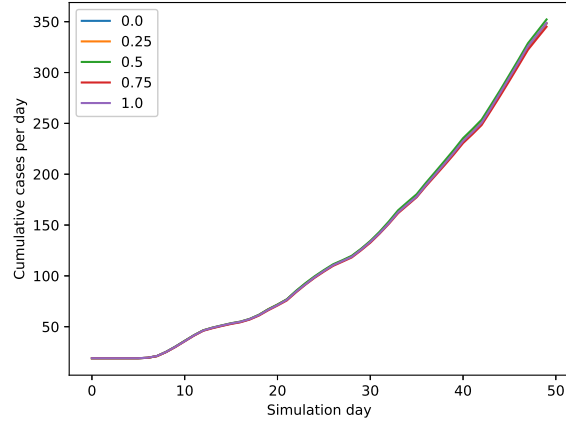
Fig. 11: Plots showing the impact of vaccinating students. Averages from 20 simulations.

### 3.3 Commuting to work

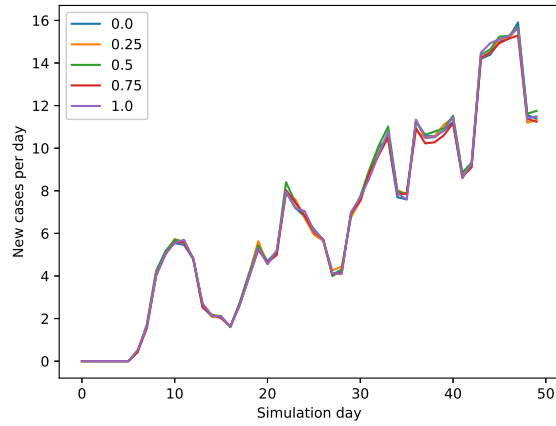
Another factor that is interesting to examine is the impact of commuting on disease spread.

In this simulation, we generate different populations, depending on different fractions of commuters. The expectation we have is that if more people commute to work, more people get infected in the end. But when we run the simulation, we see that there isn't a difference between the number of infected people when we run with different fractions of commuting. We can also see that there is a peak between day 40 and 50 (Figure 12). This peak doesn't change much when we increase the amount of commuters. If we increase the amount of days to 200 (Figure 13), the total peak is around 70-75 days, but still no difference between the number of infected people for the different fractions of commuters.

When we increase the population size from 10000 to 600000 (Figure 14), we don't see a differences in the shape of the graph. This means that the population size has no impact on the spread of this disease for this simulation.



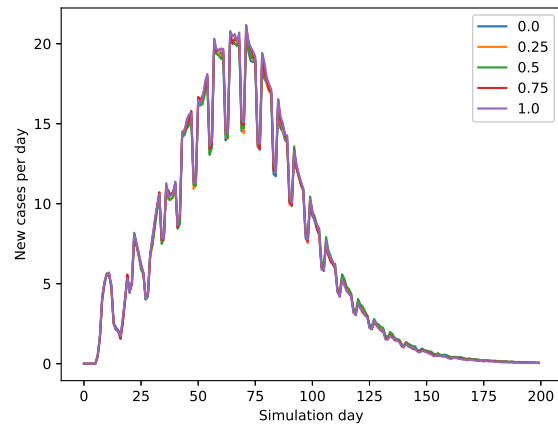
(a) Cumulative cases per day



(b) New cases per day.

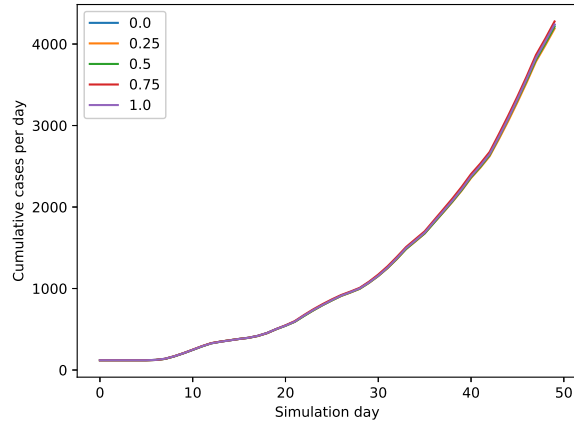
Fig. 12: Plots showing the impact of Commuting to work on a population of size 10000. 1000 simulation runs for each fraction (so 5000 in total).



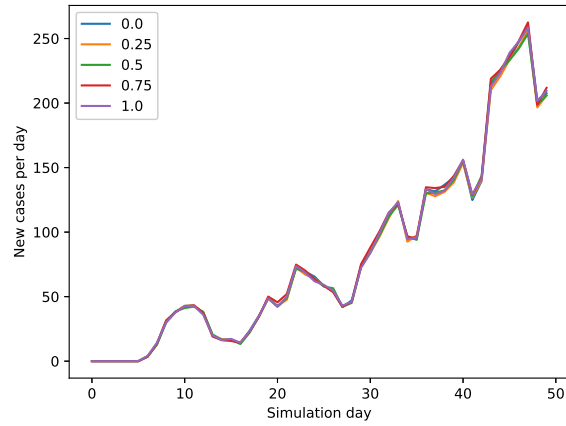


(a) New cases per day.

Fig. 13: Plots showing the impact of Commuting to work on a population of size 10000. 1000 simulation runs for each fraction (so 5000 in total). 200 days.



(a) Cumulative cases per day



(b) New cases per day.

Fig. 14: Plots showing the impact of Commuting to work on a population of size 600000. 100 simulation runs for each fraction (so 500 in total).

## 4 Performance profiling

Using GProf, we will look at the performance in different scenarios. By varying parameters, we try to see which parts of the code they have an influence on, and which parts take up the most time.

For the first 4 parameters

- amount of days
- population size
- immunity rate
- seeding rate

the actual sorting and analyzing of the population takes up most of the time.

### 4.1 Number of days

By increasing the number of days to be simulated, the total execution time gets longer as well. This should be expected as more days means more times we simulate what goes on in a day.

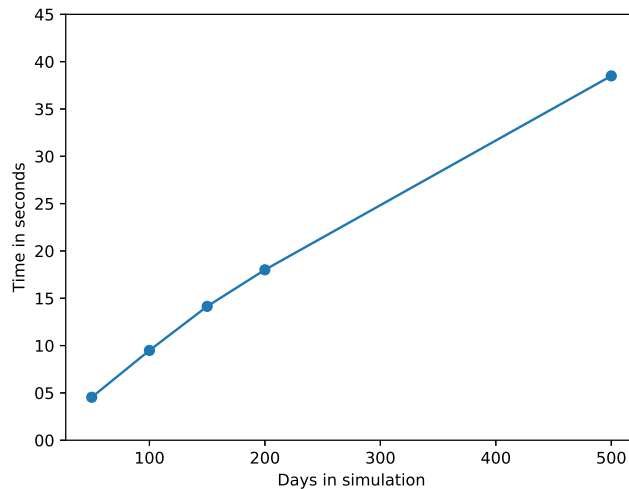


Fig. 15: Plot showing run time of simulations by varying the number of days.

### 4.2 Population size

Generating a new population depends on the given size, which was expected. The generation however is very fast.

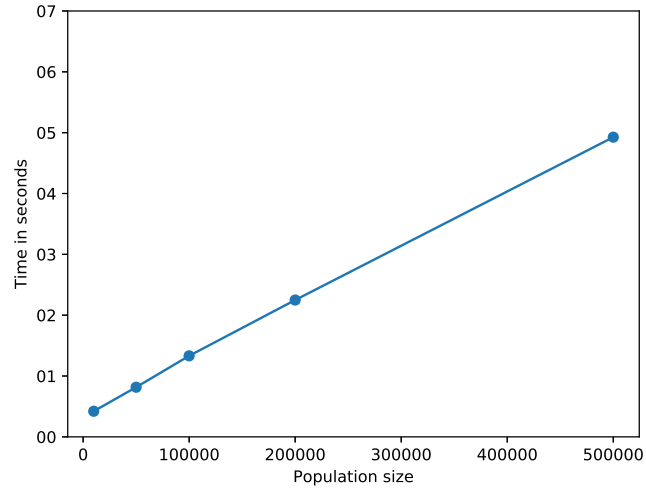


Fig. 16: Plot showing run time of simulations by varying the size of the population.

### 4.3 Immunity rate

Varying the immunity rate does not seem to affect the total execution time.

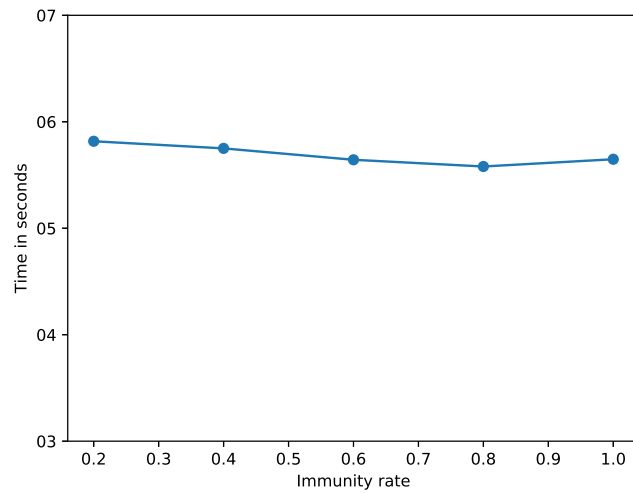


Fig. 17: Plot showing run time of simulations by varying the immunity rate.

#### 4.4 Seeding rate

A bigger seeding rate slightly increases the time of the execution. This happens as more people have to be initially infected.

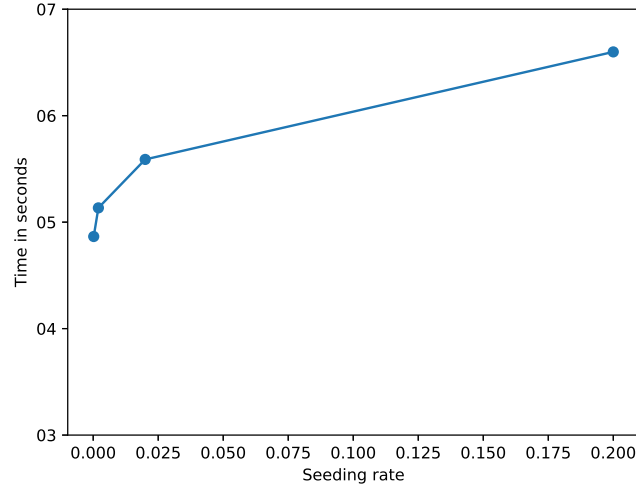


Fig. 18: Plot showing run time of simulations by varying the seeding rate.

#### 4.5 Contact log mode

The mode of the contact log has a very large impact on the execution time. Logging every contact between people takes a long time. At day 50 in the simulation, only 20000 people out of 600000 were infected. When logging the susceptible people, you actually log almost 580000 people at each day which is very close to logging all people. This is very fast when the mode is set to 'Transmissions' as you would only log once for each newly infected person. The logging modes 'All' and 'Susceptibles' use a less efficient algorithm, hence the big difference in run time between them and the modes 'Transmissions' and 'None'.

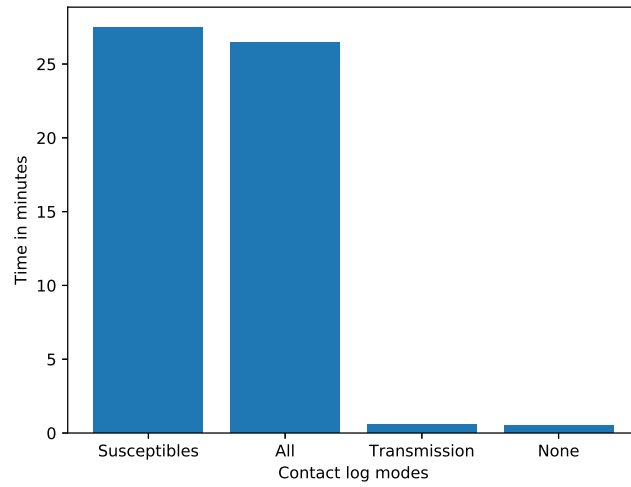


Fig. 19: Plot showing run time of simulations using different logging methods.