Bachelor Dissertation

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

University of Antwerp

Abstract. In this document we will examine the impact the changes to the Stride simulator have on the actual simulations. This is done by simulating various set scenarios in the old version of Stride as well as the updated one, and compare the results. The general aim was to achieve higher simulation realism through adding more detail to the simulator.

Keywords: Computational Epidemiology · Dissertation · Impact Paper

1 Introduction

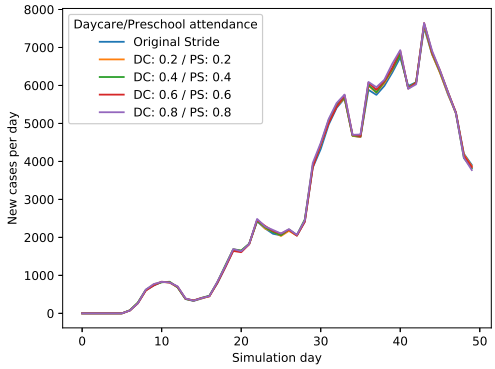
The Stride simulator was expanded upon in various aspects, detailed in the sections below. Generally, the changes did not introduce massive differences in simulation results. This was largely to be expected, as the functionality of the simulator has not been radically changed.

2 Daycare & PreSchool

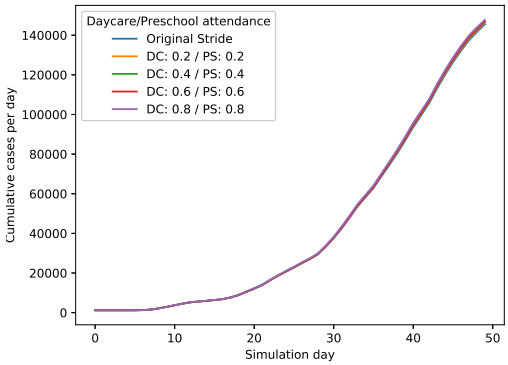
The first feature is the addition of two new ContactTypes, namely the Daycare and PreSchool. Children from ages 0 to 3 can attend Daycare, whereas children from ages 3 to 6 can go to a PreSchool. Both ContactTypes make use of a participation parameter to determine how many of these children actually will go to these ContactTypes.
The scenario used to test the impact of this addition is a simulation of the measles disease using $R_0 = 15$, a seeding rate of 0.2% and an immunity rate of 70%, in accordance with results found in the simulation paper. We compare the result of the original Stride (i.e. no daycare and preschool ContactTypes) to results found by introducing varying degrees of attendance for DayCare and PreSchool. The results are visualised in Figure 1. From this figure no significant impact on the simulation can be deduced. Further testing using a fixed attendancy for DayCare while varying PreSchool attendance or vice versa yields no different result. This is likely because despite a potentially high degree of attendance, the fraction of the population that qualifies for these ContactTypes (persons between ages 0 and 6) is too small to have a significant impact on the simulation results.

3 Demographic profile

The simulator originally only makes use of one household sample. When doing simulations for Flanders, each province and city will have the same ratio of young to old people. This is not indicative of the real ratio in each of these locations.



(a) Average new cases per day.



(b) Average cumulative cases.

Fig. 1: Simulation results using different degrees of Daycare/Preschool attendance. Averages over 20 runs, seeding rate = 0.2%, immunity rate = 70%, R_0 = 15. Simulated for a population of 600,000 over 50 days.

The algorithm for creating populations is expanded to allow the use of different household samples. Sample files have to be defined for each province, whereas they are not mandatory for the cities. When for a given city, there is no existing household sample file the algorithm will use the file defined for the entire province.

The simulations here have been done using different demographic profiles. The first profile is represents the original algorithm where only one sample file is used. The second profile uses a sample file for each province in Flanders and a file for each city that is defined as a *centrumstad*. These cities are ones with a relatively large population compared to their surroundings with a central function in the area of employment, care, education, culture and entertainment.

By visually inspecting figure 2 we can see a clear difference in the amount of infected people at the end of the simulation. The outbreaks are larger when using only 1 sample file.

We determined a 95% confidence interval for the outbreak size for the demographic coming from 1 file:

$$[97673.912, 97714.508]$$

And for the demographic coming from multiple files:

$$[95258.962, 95302.443]$$

These intervals are disjoint and lie apart quite significantly, we can conclude that using multiple household samples instead of only one has an impact on the simulation results.

4 Workplace Size Distribution

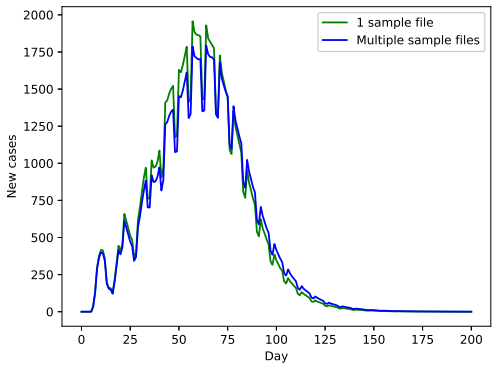
With the addition of the third feature, we can specify different ranges of sizes for workplaces. This means we can have 70% of workplaces that have a size between 20 and 50 and 30% between 50 and 100.

The workplace size distribution used in the new algorithm as the default case is the following:

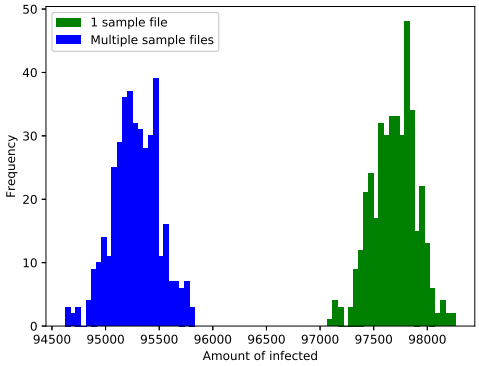
Ratio	Minimum size	Maximum size
77.854%	1	9
17.190%	10	49
4.100%	50	199
0.856%	200	400

while in the original algorithm, all the workplaces are around 20 people large.

To determine whether different size classes in the workplaces have a significant impact on the simulations, the influenza virus is used. Here, more interesting



(a) Average new cases per day.

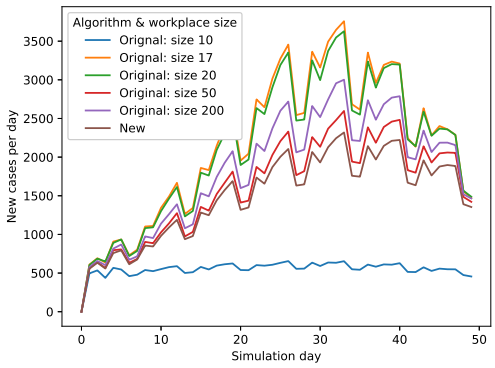


(b) Amount of infected people at the end of the simulation.

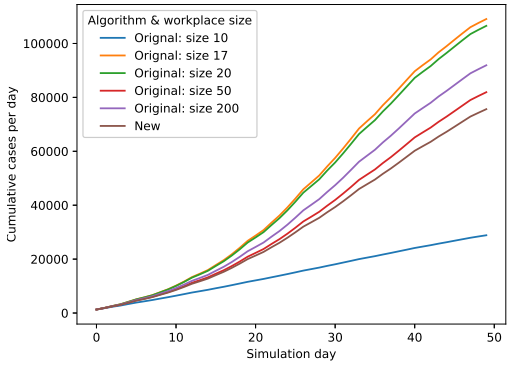
Fig. 2: Results of 400 different simulations for the 2 different demographic profiles. Seeding rate = 0.2%, immunity rate = 80%, $R_0 = 11$. Simulated for a population of 600,000 over 200 days.

results, meaning there is a noticeable difference in simulation results, can be seen when the immunity rate is sufficiently low. For this reason we used an immunity rate of 40%, this means that 60% of the population can get infected by the influenza virus. The new algorithm is compared to different sizes of workplaces using the different algorithm, with size 17 being the average for the values used in the new algorithm.

The most noticeable result is how a decreasing workplace size leads to more new cases per day, this happens until around size 10 where the amount of new cases drops significantly. Very small workplaces do not allow the disease to spread



(a) Average new cases per day.



(b) Average cumulative cases.

Fig. 3: Results of 20 different simulations for varying workplace sizes. Seeding rate = 0.2%, immunity rate = 40%, $R_0 = 2$. Simulated for a population of 600,000 over 50 days.

quickly when compared to the larger ones. A possible explanation here is that the chance to meet an infected person in these workplaces is small.

The average size of workplaces in the new algorithm does not seem to be an important factor.

While running a simulation using the original algorithm and having results of the new algorithm seems possible, finding the correct size of the workplaces is difficult. Each different size of the workplace has their own impact on the spread of the disease, so we can conclude that a combination of those sizes result in a more realistic spread.

5 Data Formats

A GeoGrid can be written to and read from three types of data formats: json, hdf5 and protobuf. Although these three formats are used for the same purpose, are there major differences in storrage capacity and between the time it takes to read from and write to a file.
To compare those times and storage capacity, we generate a default GeoGrid, write it out and read it back in.

	Write time	Read time	Storage capacity
Protobuf	7.427s	2.021s	14 MB
HDF5	34.515s	23.370s	159.8 MB
Json	23.238s	7.946s	119.2 MB

It is clearly visible that Protobuf is the overall best dataformat, followed by Json. HDF5 is the worst of the three.