VARIANCE AND SCALING IN AGENT BASED MODELLING AND SYSTEM DYNAMICS IN THE CONTEXT OF INFECTIOUS DISEASE

Aslam Ahmed, BSc(H), MSc

Thesis submitted to the University of Nottingham for the degree of
Doctor of Philosophy

Supervisors: Professor Uwe Aickelin, Dr Julie Greensmith

September 2015
Abstract

Classical approaches to epidemic simulation, such as those using System Dynamics, use differential equations to represent relationships between the dynamics of a population. The output from these simulations is obtained by solving differential equations for a given set of initial conditions over time. The behaviour of one individual represented in the population is assumed to be homogenous. For any given set of input values, a corresponding set of output curves are observed. In epidemic simulation, it can be useful to capture a range of output curves representing the underlying natural variation generated by the interaction between individuals in a population. This variation can provide a useful indicator of the range of values for which peak infection rates occur and for the range of values for the recovery time of infection. One way to do this in System Dynamics is to run repeated simulations using Monte-Carlo methods to determine input parameters for the simulation. In this instance however, the output curves are not generated as a result of the natural interaction from within members of a population but from the modification of the parameters for the simulation. Agent Based Models are able to capture natural variation in the output space without modifying the input parameters. However, with large population sizes, compared to simulations using System Dynamics models, simulations using Agent Based Model use more computational power and in some cases the performance means that the simulations are no longer feasible.

In this thesis we compare the influence and effect of variation between System Dynamics and Agent Based Modelling through a series of experiments to show that the Agent Based simulation of the epidemiological SIR model is more effective at capturing the natural variation within SIR compared to an equivalent model using
System Dynamics with Monte-Carlo simulation. Additional we show how the issue of performance with large population sizes can be addressed by scaling the population down using a combination of environment changes and the introduction of a time delay. A Susceptible-Infectious-Recovered (SIR) model is implemented using both System Dynamics with Monte-Carlo simulation and Agent Based Modelling based on previously published empirical data. In the first experiment, variation from the Agent Based Model is compared to the System Dynamics Model with Monte Carlo simulation. In the second experiment, the variation is observed across a range of input parameters. The population size is a special parameter is used to understand the effect of scaling for Agent Based Modelling. The final experiment compares the variation of Agent Based Modelling to real variation using real SIR data.
Acknowledgements

I would like to express the deepest appreciation to my supervisors, Professor Uwe Aickelin and Dr Julie Greensmith who have supported and mentored me over several years in the pursuit of an area of interest I found fascinating many years ago.

There are many people I have met and worked with over several years in the Intelligent Modelling and Analysis team including those in the weekly seminars. The seminars have been really useful and it has been a pleasure presenting the study and working with everyone.

My sincere gratitude goes to my mother and father for encouraging me and guiding me in the pursuit of learning.

My thanks go to my wife Mina and my children Ubayd and Layla. Without their continuous support and patience this would not have been possible.

Last but not least, I would like to thank my older brother Mahmood who has motivated me from an early age and supported me through the many years of my studies.
## Contents

1. Introduction ....................................................................................................... 1
   1.1. Background and Motivation ........................................................................ 1
   1.2. Research Aims and Objectives ..................................................................... 4
   1.3. Thesis Contribution ..................................................................................... 5
   1.4. Published Papers ......................................................................................... 6
   1.5. Experiments and Findings ........................................................................... 7
   1.6. Summary of Thesis Outline ......................................................................... 7

2. Literature Review ............................................................................................ 10
   2.1. Introduction ............................................................................................... 10
   2.2. Modelling of SIR ....................................................................................... 10
   2.3. Computer Simulation Modelling Paradigms ................................................ 12
   2.4. Agent Based Modelling ............................................................................. 14
   2.5. System Dynamics ...................................................................................... 33
   2.6. Related Studies of Variance using ABM and SD ....................................... 45
   2.7. Adding randomness to System Dynamics .................................................. 52
   2.8. Conclusions ............................................................................................... 53

3. Initial Experiments using the SIR Model .......................................................... 58
   3.1. Introduction ............................................................................................... 58
   3.2. System Dynamics Model of SIR ............................................................... 61
   3.3. Agent Based Model of SIR ........................................................................ 62
3.4. Variance of ABM compared to SD using the SIR Model ......................... 68
3.5. Variance of ABM using the Population Parameter for the SIR Model ....... 81
3.6. Conclusions ............................................................................................... 96

4. Variation of ABM and SD across Stepped Parameter Ranges ...................... 98
4.1. Introduction ............................................................................................... 98
4.2. Variation of ABM and SD for SIR ............................................................. 99
4.3. Conclusions ............................................................................................. 125

5. Variation of ABM Output Compared to Real Variation .................................. 127
5.1. Introduction ............................................................................................. 127
5.2. Results for Variation of ABM Compared to Real Data .............................. 141
5.3. Analysis of Results for Variation of ABM Compared to Real Data .......... 157
5.4. Conclusions ............................................................................................. 163

6. Conclusions ................................................................................................... 170
6.1. Introduction ............................................................................................. 170
6.2. Research Findings ................................................................................... 171
6.3. Contribution to Knowledge ...................................................................... 173
6.4. Limitations .............................................................................................. 177
6.5. Further work ............................................................................................ 179

7. References ..................................................................................................... 181
List of Figures

Figure 1: Neighbourhoods for Cellular Automata .................................................... 16
Figure 2: Common network topologies ................................................................. 18
Figure 3: A definition of an Agent in an Agent Based Model ............................ 24
Figure 4: A definition of the states of Agents in an Agent Based Model of SIR .... 25
Figure 5: Update method for the ABM from 'To Agent Based Simulation from System Dynamics' [60] ........................................................................................................ 26
Figure 6: The different profiles of SD Output [66] ................................................... 36
Figure 7: Causal loop diagram for births and deaths ............................................. 37
Figure 8: Stock and flow diagram for births and deaths.......................................... 37
Figure 9: An example of how ABM Simulation output varies across two simulations ......................................................................................................................... 46
Figure 10: A comparison of different modelling solutions for SIR [60]................. 48
Figure 11: A view of the variance across one thousand simulations for SIR using Agent Based Modelling [60] ........................................................................................................ 49
Figure 12: Output for the SEIR model using the small world network from Rahmandad and Sterman [86] ..................................................................................... 50
Figure 13: Output for the SEIR model using the scale free network from Rahmandad and Sterman [86] .............................................................................. 51
Figure 14: Infected population values for the region of Österlövsta in the Russian influenza in Sweden based from Skog and Linde [96] ................................. 60
Figure 15: A Stocks and Flows diagram for the System Dynamics Model for SIR ... 62
Figure 16: Results for the System Dynamics Model for SIR ............................... 62
Figure 17: State chart for the Agent Based Model for SIR ................................. 64
Figure 18: Results of the Agent Based Model for SIR .............................................. 67
Figure 19: System Dynamics Model for SIR with Illness Duration Variation ......... 71
Figure 20: System Dynamics Model for SIR with Contact Rate Variation .......... 72
Figure 21: System Dynamics Model for SIR with Infection Rate Variation .......... 72
Figure 22: System Dynamics Model for SIR with Infection, Illness and Contact Rate Variation .......................................................... 73
Figure 23: Interquartile range for the Agent Based Model for SIR ...................... 75
Figure 24: Interquartile range for the System Dynamics Model for SIR with Illness Duration Variation ................................................................. 76
Figure 25: Interquartile range for the System Dynamics Model for SIR with Contact Rate Variation ................................................................. 77
Figure 26 Interquartile range for the System Dynamics Model with Infection Rate Variation ................................................................. 78
Figure 27: Interquartile range for the System Dynamics Model with Infection, Illness and Contact Rate Variation ................................................................. 78
Figure 28: Output from the Initial Agent Based Model for SIR ......................... 87
Figure 29: Infected population values for the region of Österlövsta in the Russian influenza in Sweden based from Skog and Linde, 2008 ......................... 88
Figure 30: Output from the initial Agent Based Model with Population size 5291 . 89
Figure 31: Output from the Initial model with population size 529 ..................... 90
Figure 32: Output from the Agent Based Model using updated environment with population size 5291 ................................................................. 91
Figure 33: Output from the Agent Based Model using updated environment with population size 529 ................................................................. 92
Figure 34: Output from the updated environment using delay with population size 5291 ........................................................................................................................ 93
Figure 35: Output from the updated environment using delay with population size 529 ................................................................................................................................ 94
Figure 36: Overview of the process for the tipping point experiments ...................... 99
Figure 37: Output for the Agent Based Model for SIR with a Population of 52910 102
Figure 38: Output from the SD model for SIR with Population of 52910 and Illness Duration Variation ................................................................................................. 105
Figure 39: Output from the ABM model with Population of 52910 and Illness Duration Variation ................................................................................................. 105
Figure 40: Hypothesis test for the SD model with Population 52910 and Illness Duration Variation ................................................................................................. 106
Figure 41: Hypothesis test for ABM with Population 52910 and Illness duration Variation ................................................................................................. 106
Figure 42: Detailed hypothesis test for SD model with Population 52910 and Illness Duration Variation ................................................................................................. 107
Figure 43: Detailed hypothesis test for ABM with Population 52910 and Illness Duration Variation ................................................................................................. 107
Figure 44: Output from SD model with Population 5291 and Illness Duration Variation ........................................................................................................................ 108
Figure 45: Output from ABM with Population 5291 and Illness Duration Variation ........................................................................................................................ 108
Figure 46: Hypothesis test for SD model with Population 5291 and Illness Duration Variation ........................................................................................................................ 109
Figure 47: Hypothesis test for ABM with Population 5291 and Illness duration Variation .................................................................................................................. 109

Figure 48: Detailed hypothesis test for SD model with Population 5291 and Illness Duration Variation .................................................................................................................. 110

Figure 49: Detailed hypothesis test for ABM with Population 5291 and Illness Duration Variation .................................................................................................................. 110

Figure 50: Output from SD model with Population 529 and Illness Duration Variation .............................................................................................................................. 111

Figure 51: Output from ABM with Population 529 and Illness Duration Variation .............................................................................................................................. 111

Figure 52: Hypothesis test for SD model with Population 529 and Illness Duration Variation .................................................................................................................. 112

Figure 53: Hypothesis test for ABM with Population 529 and Illness Duration Variation .................................................................................................................. 112

Figure 54: Detailed hypothesis test for SD model with Population 529 and Illness Duration Variation .................................................................................................................. 113

Figure 55: Detailed hypothesis test for ABM with Population 529 and Illness Duration Variation .................................................................................................................. 113

Figure 56: Output from SD model with Population 52910 and Contact Rate Variation .............................................................................................................................. 114

Figure 57: Output from ABM with Population 52910 and Contact Rate Variation .............................................................................................................................. 114

Figure 58: Hypothesis test for SD model with Population 52910 and Contact Rate Variation .................................................................................................................. 115

Figure 59: Hypothesis test for ABM with Population 52910 and Contact Rate Variation .................................................................................................................. 115
Figure 60: Detailed hypothesis test for SD model with Population 52910 and Contact Rate Variation ................................................................. 116
Figure 61: Detailed hypothesis test for ABM with Population 52910 and Contact Rate Variation ........................................................................................................... 116
Figure 62: Output from SD model with Population 5291 and Contact Rate Variation .......................................................................................................................... 117
Figure 63: Output from ABM with Population 5291 and Contact Rate Variation ............................................................................................................................. 117
Figure 64: Hypothesis test for SD model with Population 5291 and Contact Rate Variation ...................................................................................................................... 117
Figure 65: Hypothesis test for ABM with Population 5291 and Contact Rate Variation ............................................................................................................................. 118
Figure 66: Detailed hypothesis test for SD model with Population 5291 and Contact Rate Variation ...................................................................................................................... 118
Figure 67: Detailed hypothesis test for ABM with Population 5291 and Contact Rate Variation ...................................................................................................................... 118
Figure 68: Output for SD model with Population 529 and Contact Rate Variation . 119
Figure 69: Output for ABM with Population 529 and Contact Rate Variation ...... 119
Figure 70: Hypothesis test for SD model with Population 529 and Contact Rate Variation ............................................................................................................................. 120
Figure 71: Hypothesis test for ABM with Population 529 and Contact Rate Variation ............................................................................................................................. 120
Figure 72: Detailed hypothesis test for SD model with Population 529 and Contact Rate Variation ...................................................................................................................... 121
Figure 73: Detailed hypothesis test for ABM with Population 529 and Contact Rate Variation ...................................................................................................................... 121
Figure 74: United States Census Divisions from U.S. Energy Information Administration, Office of Energy Analysis ................................................................. 129
Figure 75: Influenza data for 2008-2009 ................................................................. 131
Figure 76: Influenza data for 2009-2010 ................................................................. 132
Figure 77: Influenza data for 2010-2011 ................................................................. 133
Figure 78: Flue data for 2011-2012 ...................................................................... 134
Figure 79: Influenza data for 2012-2013 ................................................................. 135
Figure 80: Aggregate Death Rates for US Regions for 1918-1919 Influenza Pandemic from [139] ................................................................................................. 135
Figure 81: Influenza data for 2013-2014 ................................................................. 136
Figure 82: Agent Based Model for analysing influenza variation ......................... 137
Figure 83: Strategy for determining model parameters .......................................... 139
Figure 84: Infected population for New England 2008-2014 .............................. 144
Figure 85: Infected population using an Agent Based Model with 6 runs .......... 145
Figure 86: Infected population using an Agent Based Model with 1000 runs ....... 146
Figure 87: Median values from influenza data from CDC and Agent Based Model 147
Figure 88: Interquartile ranges for CDC data and ABM data ............................... 148
Figure 89: Infected Population as connections per agent are varied ................. 150
Figure 90: Interquartile range as connections per agent are varied ..................... 152
Figure 91: Infected Population as neighbour link fraction is varied ................... 154
Figure 92: Interquartile range as neighbourhood link fraction is varied .............. 156
Figure 93: Representation of local clustering in an Agent Based Model .............. 162
List of Tables

Table 1: The different types of links in System Dynamics.............................. 35
Table 2: A Legend for the different elements shown in the output charts for the Agent Based and System Dynamics models from Rahmandad and Sterman [86]............. 50
Table 3 : System Dynamics compared to Agent Based Modelling from Denton [89] 54
Table 4: A description of the parameters for the SIR model ................................. 61
Table 5: Wilcoxon Rank Sum Test for the SIR variance experiments ....................... 74
Table 6: Total Variation observed for the SIR Experiments ................................. 79
Table 7: Specification of the computer used to perform the scaling experiments ...... 84
Table 8: Time required for running stepping experiments for a population size of 529,100 ................................................................................................................. 103
Table 9: Total variation for Illness Duration for Agent Based Model of SIR ........ 113
Table 10: Total variation for Contact Rate for Agent Based Model of SIR ............. 121
Table 11: Influenza variants from CDC data for 2008 to 2014 ............................... 130
Table 12: Network parameters for the Agent Based Model ................................. 137
Table 13: Input parameters for the Agent Based Model......................................... 141
Table 14: Median Statistics for CDC and ABM curves ....................................... 147
Table 15: Sum of interquartile range as connections are varied ......................... 151
Table 16: Sum of interquartile range as neighbour link fraction is varied ............... 157
List of Abbreviations

ABM - Agent Based Modelling

SDM - System Dynamics Modelling

DES - Discrete Event Simulation

SIR - Susceptible, Infectious, Recovered

SEIR - Susceptible, Exposed, Infectious and Recovered

CDC - Centres for Disease Control

ODD - Overview, Design concepts and Details

IQR - Interquartile Range
1. Introduction

1.1. Background and Motivation

Computer simulations are abstractions of the real world implemented in a computer system that can be used to make decisions and increase our understanding of how the complex real world systems bring about observed behaviour [1]. They are of major significance across a whole series of different areas of work used for monitoring and helping to understand a range of systems at the molecular level to the behaviour of whole societies and used as key influencers of decision making in an ever-increasing world of technology and automation [2].

Traditional methods of understanding or gaining insight of the real world are based on a series of ordinary differential equations which are usually coupled. An ordinary differential equation involves an unknown function of a single variable, its dependent variable and one or more of its derivatives [3].

Part of the popularity for these type of methods comes about from well understood mathematical laws of nature which involve how the rate of one variable changes with respect to another [4] such as the laws of motion and thermodynamics. Gaining insight of the real world using these methods involves solving the differential equations based on a set of specified input parameters. They are quick to solve, based on well understood foundations and can be done on paper or computer.

For simulations based on differential equations, if the simulation is re-run using the same input parameters, the same set of output results or output curve is generated so there is a direct one-to-one mapping between the input parameters and the output
from the simulation. In many cases, this may be appropriate depending on the nature of the problem being solved or the nature of the research being carried out [5]. Simulations are abstractions of the real world and if the simulation is easy to repeat, quick to run and offers a clear result from which inference can be drawn, then it can be considered that the simulation has met the needs to the study.

However, depending on the nature of an investigation, it may be desirable for a simulation to supply a broad range of outcomes taking into account knowledge of the real world in order to reduce risk and uncertainty [6]. Real world systems are highly complex and generally non-deterministic. Small variations in the behaviour of real world systems can become exaggerated and lead to different outcomes and there is often a need to evaluate simulation performance in the face of uncertainty and to develop models that faithfully represent the sources of uncertainty in the systems they simulate [7].

In the case of influenza for example, an outbreak may occur in a small localised area but depending on the nature of contact among individuals in a population, it can spread rapidly [8]. The extent of the spread can be attributed to the strength of the connections between different individuals based on locality. In such scenarios, it is useful to understand the range of times for which peak infection may occur so that suitable medical resources and facilities may be arranged during those times rather than having a single definitive value for peak levels of infection [9]. This can be more useful than a single value for the time for peak infection.

We consider two main paradigms of computer simulation: System Dynamics Modelling (SDM), and Agent Based Modelling (ABM). SDM examines how the
different components of a system change over time [71]. ABM examines systems in terms of interactions between autonomous components of the system over time [10]. These autonomous components are known as 'agents'.

In this thesis, the spread of results in the output space (which represents the range of output values for the simulation over time) for the different simulation paradigms is examined through a series of experiments and referred to as 'variation' in the output space for the simulations. Three sets of experiments are carried out to show how:

1. The variation of ABM without input parameter modifications compares to SDM with Monte Carlo simulation.

2. The variation of ABM compares to SDM with input parameters that are stepped across a defined range for both models. The 'stepping' refers to the increment applied to a parameter from the minimum value in the range to the maximum value.

   One special parameter that is varied is the population size, which is used to understand the impact of scaling for ABM.

3. The variation of ABM compares to the variation in real data.

Classic forms of mathematical simulation such as those used in SDM are based on ordinary differential equations and are considered to have no direct variance in the output space. One way to introduce variation is to use Monte-Carlo methods [11] to determine the input parameters for the simulation. As the input parameters are probabilistic values, the output curves vary for each simulation.

One of the main issues encountered with ABM is that of performance, or specifically the time taken to complete an experiment comprising many individual simulations.
1. Introduction

The individual simulations are necessary in most cases as the simulation output is different on each run and therefore a number of simulations are required in order to obtain a mean or median set of output values to validate the model. Traditionally, this has made ABM infeasible due to limited computing resource but recent advances in computer performance coupled with the availability of off-the-shelf ABM packages such as AnyLogic [12] has made ABM more popular.

However, with large population sizes comprising hundreds of thousands or millions of agents in complex environments, compared to simulations using SDM models, simulations using ABM run much slower and in some cases the performance means that the simulation is no longer feasible [13] on standard desktop computing platforms.

1.2. Research Aims and Objectives

The aim of this thesis is to investigate the influence and effect of variation and scaling between SDM and ABM, through a series of experiments.

In the initial set of experiments described in chapter three, the aim is to compare the variation generated by an ABM Model without any input parameter modification to an SD with Monte Carlo simulation. An SD model and ABM of the epidemiological SIR (Susceptible, Infectious, and Recovered) model is implemented and validated against previously published empirical data.

In the second set of experiments in chapter four, incremental changes are made to the different input parameters of the SIR model to determine the effect of the changes on the output space of the simulation for both SDM and ABM. During each change, a Wilcoxon rank sum test is applied to the output curve of the simulation against the
original experimental data to map the space of the parameters which is in agreement or disagreement with the original experimental data. The population size is a special parameter that is used in these experiments to determine the scalability of ABM with respect to population size.

The aim of the final set of experiments in chapter five is to determine how variation from ABM compares to that of real data.

1.3. Thesis Contribution

The main contribution of this thesis is to provide a detailed study of how the variation in the output space generated by ABM compares to the output space from SDM. The second contribution of this thesis is an examination of scaling of different sizes for the number of agents for ABM.

ABM has grown steadily alongside interest in complex systems science and advances in technology. What isn't currently well understood in the field of ABM is the variation generated by a simulation from ABM and how realistic this variation is compared to real variation.

For the main contribution of variation, this research examines the variation of the output curves generated in the output space for ABM described in chapters three and four. This is done through a series of experiments which examines the profile of the variation across repeated simulations, environment changes and input parameter modification. The main model used for this study is the Susceptible Infectious Recovered model using empirical data. This shows how the variation of ABM compares to the variation from SDM across the same set of input parameters.
There is currently very limited research where the variation in the output space of ABM is compared to variation in real world systems. This is covered in Chapter five where the variation of ABM is compared to the variation of a real world system.

The second contribution is in the area of scaling for ABM. Real world systems vary in terms of the number of agents. These are typically population sizes in social sciences and in models of the transmission of infectious disease.

Population sizes depend on the nature of the system being modelled but these can run into many tens of millions or hundreds of millions or more depending on the geographic area in the real world being modelled. This makes ABM less attractive as an option for computer simulation as the simulation time can exceed the boundaries of acceptable time constraints.

The scaling study in chapter four examines the effect of changes in the population size in ABM using the Susceptible Infectious Recovered Model in terms of speed, variation and the profile of the output curve for the simulation.

1.4. Published Papers

The following publication has been produced as part of this thesis:


This paper was nominated for the best paper award for the conference. For this paper, the major contribution was made by Aslam Ahmed.
1.5. Experiments and Findings

For the experiments on variation in the output space of ABM, it is found that the variation generated by an ABM Model can be tailored to match the variation of an equivalent real world system as determined by a Wilcoxon Rank Sum test [14] using the averages of the output for the ABM Model across time. For the matching of the variation to take place, both the input parameters for the model and environment configurations needs to be considered.

For the experiments on scaling, it is found that it is possible to generate equivalent ABM Models across several population sizes with similar levels of variation across each population size. One aspect of the output curve found during the scaling is the requirement of a time delay to line up the output curves for the model so that the peak values for the curves occur at the same time value.

Based on the experiments, we establish that ABM can provide a range of output curves which when analysed, can be used to gain additional insight into the range of peak levels of output possibilities. This is very useful for the modelling of infectious diseases where medical resources are finite and they need to be distributed in the best manner to reduce the total infectivity of a population based on knowledge of the location of peak infection values across time.

1.6. Summary of Thesis Outline

Chapter one provides an overview of the motivation, aims and objectives of the research which is based on the variation and scaling aspects of ABM relative to SDM.

Chapter two describes the research background and literature review. The chapter is split into different sections. The first provides an introduction to computer modelling
and simulation. The second section discusses the different paradigms available for computer modelling and simulation. The third, fourth and fifth sections describe these paradigms under the sections of ABM and SDM. The sixth section describes related work in the field of variance using ABM and SDM. The final section of the chapter, section seven, provides a summary for the background research and literature review.

Chapter three covers the initial set of experiments including the experiment design and describes the main model, the SIR (Susceptible, Infectious and Recovered) Model which is used to perform the variance and scaling experiments. The first section in the chapter starts with an introduction of the origins of the Susceptible, Infectious, Recovered Model. Sections two and three show how this is implemented in SDM and ABM respectively. The models are validated using empirical data by using a Wilcoxon Rank Sum test against the output curves. Two sets of experiments are performed. In section four, the first experiment is done to understand how the variance in ABM compares to the variance from SDM combined with Monte Carlo simulation to select the input parameters. The input parameters for the ABM Model are not stepped in this experiment as this is a validation exercise. In section five, the second experiment covers the effect of the population parameter which is a special parameter used to understand scaling aspects of ABM. The chapter ends with section six which contains the conclusions including a description of the results and how the model is used in the experiments.

Chapter four examines how the variance of ABM compares to SDM when parameter variation is applied to both models. Instead of choosing Monte Carlo simulation to select the input parameters, the input parameters are incremented across a range to provide consistency for comparison. The chapter starts with an introduction in section
one and this is followed by the experiments in section two which covers the research into the effect of stepping the input parameters using the SIR model. Section three contains the conclusion with a summary of the results.

Chapter five examines how the output variation from ABM compares to that of variation in the real world. There is an introduction to the characteristics of the data in section one followed by a description of the model and the results are presented in section two. Section three provides an analysis of the results and a conclusion is contained in section four with a summary of the findings.

The thesis concludes with chapter six which is split into five sections. The first section provides a summary of the work. The second section undertakes a review of the findings. The third section looks at the contribution of the work towards current knowledge in the field. Section four discusses some of the limitations of the work. This is followed up by section five which provides a discussion of further work to take the research forward and to try to address the limitations found.
2. Literature Review

2.1. Introduction

Computer simulation is the use of a computer or computers to 'simulate' the operation of a real world process which is usually termed a 'system' [15]. To generate a scientific study, a set of assumptions are required and these constitute what is known as a 'model' which can be based on mathematical methods or logical relationships.

Models whose components are simple can be based on mathematical methods such as algebra or calculus and completed manually using paper-based methods. These are termed 'analytical' solutions. Many real-world systems are too complex for analytical solutions and require simulations to evaluate models numerically to make estimations for the characteristics of the model such as those in finance [16] and engineering [17].

This chapter introduces the main groups of modelling paradigms for computer simulation. Following this, the SIR model is presented as this is the model used in the experiments in this thesis. Next the origins, uses, strengths and limitations of ABM [18] and SDM [19] as modelling paradigms is discussed. ABM is a relatively newer modelling paradigm compared to SDM so some of the earlier related ideas are introduced. The chapter summaries the background research and concludes with a description of the research hypothesis.

2.2. Modelling of SIR

2.2.1. Introduction

The SIR model was originally proposed by Kermack and McKendrick [104] and is termed a compartmental model [24] as the population is effectively divided into three
compartments or states depending on their relationship to infectious disease. Over time, the state of an individual flows from one compartment to another.

In the basic SIR model of infectious disease used in this thesis, these compartments are Susceptible (S), Infectious (I) and Recovered (R). Susceptible individuals have never encountered an infection. Once infected, individuals move to the recovered state. The SIR model has been chosen in this thesis as it is a simple model which allows more focus to be placed on the investigation of variation. As models becomes more complex, this also has a knock-on effect of increasing the amount of qualitative data required to build and verify them, which can become an obstacle to the main investigation.

The original SIR model comprises a set of closely coupled differential equations. For this reason, the most popular method for modelling SIR involves solving these calculations over time for a given set of inputs. These equations can be solved using calculators and additionally a number of software tools exist to enable these to be solved easily. These include MatLab, Wolfram Mathematica and Vensim.

The dynamics of infectious disease is very complex and most models of infectious disease are highly abstractive of the physical nature of infections. Additional variants of SIR include:

- MSIR. This refers to Maternally derived immunity.
- Susceptible, Infectious and Susceptible. Infections such as the common cold do not result in permanent immunity, therefore, once infected, a person returns to a susceptible state.
2. Literature Review

- Susceptible, Exposed, Infectious and Recovered (SEIR). This includes an exposed state which represents an incubation time for the infection where an individual is infected but not able to pass on the infection.

2.3. Computer Simulation Modelling Paradigms

Mainstream transition of simulation from hand-written analytic methods to computer simulation came about with the availability of affordable desktop computers therefore the rise in the use of computer simulation has grown alongside physical computer inventions and technology advances [25]. Also, over time there has been a steady need to run ever more complex forms of computer simulation [26] leading to performance issues particularly at desktop level. This is in part due to the nature of the way that simulations are performed on desktop computers. Traditional desktop computers contained a single CPU such as the Intel Pentium 4, launched in 2000 with a single core executing computer commands one step at a time. Current desktop machines can have processors such as the Intel Core™ i7-3930K, launched in 2011 with 6 cores executing computer commands at the same time.

Although speeds have steadily increased, on the whole processing is still limited to a few concurrent commands at the desktop level. However a number of advances have made it possible to increase the level of concurrency at a desktop level. These include the availability of hyper-threading [27] for physical cores and the availability of computer simulation using graphical processing units available on the computer. Compute Unified Device Architecture (CUDA) [28] is a platform developed by NVIDIA to support the development of computer software using the GPU.
There are many different classifications of computer simulation which can be
categorised on whether they are stochastic or deterministic, local or distributed or the
kind of environment they are based on. Some types of computer simulation can also
contain a combination of these categories. Among these different classifications three
distinct paradigms of computer simulation can be identified:

- SDM
- ABM
- Discrete Event Simulation (DES)

In the next chapters of this thesis, SDM and ABM are reviewed. A short history of the
origins of each paradigm is presented together with a review of the applications for
each paradigm, the development tools available, the type of variance they are able to
offer and a summary of the relative performance.

The variance and performance can be key components in the choice of modelling
paradigm to use for a specific area of research so these are presented as a summary in
the literature review and discussed in more detail as part of the experiments.

This thesis compares the variation from a classical equation based method to the
variation from an individual based method. ABM has been selected for the individual
based method as there is a natural fit where the agents are representations of
individuals. Therefore the use of DES is not discussed in this thesis.
2.4. Agent Based Modelling

2.4.1. Introduction

ABM looks at the world using a bottom-up approach using local level interaction to generate outcomes [29].

What we consider now as the main components of ABM are:

- The environment
- The agents and their associated properties
- Local rules of interaction

For example, in the case of the spread of infectious disease, it is the contact from local rules of interaction that generates the spread of infection which has a cascading effect leading to what we might consider as an epidemic. The epidemic is not defined in terms of the product of the sum of the different interactions; rather it is defined as a product of the interactions themselves when allowed to continue for a fixed amount of time. They are therefore closely related to the work on complex systems which are used to describe self-organising autonomous entities whose interactions lead to emergent behaviour [30].

Emergent behaviour is behaviour of a system brought about by interactions among the components of the system that by themselves do not exhibit the behaviour. The term 'Agent' can be referred to as any individual element of the model [31].

2.4.2. Cellular Automata

The ideas of ABM and these components can be traced back to the work on cellular automata [32] in the 1940s by Stanislaw Ulam and John von Neumann. A cellular
automata comprises a grid in which each cell in the grid has one of a possible set of states (for example, on or off). All the cells in the grid have an initial state. A new state is created for the cells (termed a new generation) when each cell in the grid follows a rule based on its neighbouring cells.

In the 1970s, Conway's 'Game of Life' [33] showed how simple cellular automata obeying some simple rules is able to generate a series of stable and alternating states. This was at a time where the idea of complex systems became more accepted and what made Conway's work interesting was the idea that these stable states could not be perceived by simply analysing the rules but by running the simulation. The simulation does not necessarily need to be 'run' on a computer but doing so by hand is very time consuming.

Further, more detailed work on cellular automata was carried out by Wolfram in the 1980s which later resulted in an extensive publication with detailed analysis and examples of many different types of cellular automata [34].

There is one key aspect of cellular automata that is notable compared to the idea of using SDM and that is the idea of an environment. SDM is based on a regular, homogenous environment where there is mixing of key parameters. For example, contact between people varies in a population at individual level but this is represented as an average value in SDM. The state or output from SD as time progresses is not based on local environment events but based on the solutions generated from earlier parameters. Here, in cellular automata, was the notion of decentralised control. The state of the simulation at any given time for cellular
automata is based on a set of local decisions carried out from the initial state up to the
chosen time, leading to emergent behaviour.

2.4.3. Neighbourhoods

Neighbourhoods are well defined spatial areas where interactions take place [35].
Early environments for cellular automata were very simple. Conway's game of life
used the Moore neighbourhood [36] on a square grid. The common neighbourhoods
for cellular automata are shown in Figure 1.

![Neighbourhoods for Cellular Automata](image)

**Figure 1: Neighbourhoods for Cellular Automata**

The r value denotes the range. Based on this, a different spatial layer is generated for
the Von Neumann neighbourhood and the Moore Neighbourhood. This can be
expressed mathematically. The \( C \) value denotes the total number of squares which
make up the neighbourhood shown as dark grey squares in Figure 1.

Von Neumann:

\[
N^\nu_{(x_0, y_0)} = \{(x, y) : |x - x_0| + |y - y_0| \leq r \}
\]

\[C = 2r(r + 1) + 1\]
2. Literature Review

Moore neighbourhood:

\[ N_{(x_0,y_0)}^m = \{(x,y) : |x - x_0| \leq r, |y - y_0| \leq r \} \]

\[ C = (2r + 1)^2 \]

Although cellular automata have been used to solve a variety of problems in settings including biology [37] and traffic management [38], their simplicity means that the model can be overly abstracted. However, one of the advantages of the Moore neighbourhood is that simulations are fast to complete.

In the case of the modelling of transmission of infection, we have a complex environment and a square grid is too simplistic as it assumes that every individual falls within one of the predefined grids and the generation rules are based on the immediate locality of the locations adjacent to the cell. In the case of infectious disease, although transmission is often based on immediate locality, there is also transmission to other areas which are the result of movement of people.

ABM encompass the ideas of cellular automata in the sense of localised decision making in the form of local rules of interaction but provide a more comprehensive set of modelling tools [39].

2.4.4. Network Topologies

In computer simulation there are several well known defined topologies [40] that can be used to describe the environment of the system as shown in Figure 2.
The topologies extend the notion of a grid based model. This is usually in the form of a continuous space where the elements or agents of the system reside and interact. The common topologies include the regular, randomly connected, small world and scale free networks but custom network topologies can also be built.

The ring lattice is a simple network with 'local connections' so there is a connection from one entity to a number of closer entities. The small world network topology
proposed by Watts and Strogatz [41] supports the idea of local clustering by rewiring the connections in a regular structure based on a probability that the edge will be re-assigned. Where all rewiring takes place, we end up with a random network topology.

The small world network is a useful topology for studying infectious disease because the local clusters can be considered as a set of connected geographical areas with the connections made possible by travel from one region to another [42]. Scale free networks are not as clustered as small world networks and more hierarchical. Whilst not closely representing local clusters, these too have been used in a number of epidemic-related studies [43].

What is most useful about the small world network is that it is possible to use it to represent different levels of abstraction. For analysing the spread of infectious diseases at a national level, the agents can be considered as people and the clusters of people can be considered as cities within a country. For analysing the spread of infection at a county level, the clusters of agents can be considered as the different counties within the country. The choice of suitable abstraction level depends on the nature of the investigation. It is also possible to combine multiple small world networks in a group, each of which is a small world network in its own right [44].

One of the main limitations of the common predefined network topologies is that the movement of the agents is not explicit but represented through the connection time (sometimes considered as the length) between the transfer of infection from one individual to another. So agents which are connected by a greater distance implies that it takes longer for contact to be made from one agent to another. The actual mode of
movement from one agent to another is not considered as the physical location of the agents are not maintained.

To determine real-life networks in infectious disease is difficult and time consuming as it requires knowledge of each individual and the contact made between others [45]. This can be made even more difficult because the flagging of contact or no contact depends on the type of system being modelled. For infection in the SIR model, we need to define how we decide which people to include for the purpose of infection transmission. For example, does physical presence count as a contact or is the contact only noted when there is a chance that infection may occur? One method to account for this is to use a weighting factor within the contact but this itself can lead to difficulties as the weighting rules need to be defined, which may not be accurate.

To understand very specific scenarios, such as the effect of closing public transport on the time it takes for an infectious disease to peak, it is sometimes necessary to include a custom environment [46]. The environment in this scenario can be continuous but the agents can contain a co-ordinate based property moving along a series of links to other agents. These links can be categorised via different forms of transport which can include by foot, by car or by public transport. It would therefore be possible to simulate the effect of turning off one or more different forms of public transport on the overall spread of infection, although some level of abstraction would be required as the computational time for modelling this level of complexity on current desktop hardware for the entire population of a nation would make simulation time unfeasible.
2.4.5. The Application of Agent Based Modelling

ABM has become very popular in modelling for social sciences where agents represent people in communities. Many of these studies are based around the movement of people, often between urban and rural areas to understand the social and economic driving factors for peoples movement [47]. Others examine the cause and effect of individual behaviours such as cigarette smoking on the population as a whole [48].

The Schelling Segregation [49] model has been studied on a number of occasions with ABM. The Schelling Segregation model introduced by Schelling in 1971, shows how behaviour of individuals in a large community can lead to segregation by ethnicity. This is done by the behaviour of individuals who have a tendency to reside in an area where there are similar neighbours. Over time, this leads to distinct areas of a large community with individuals of similar ethnicity.

The Schelling model uses a rectangular space of cells and in each cell there can only be a single person (representing a resident) which can belong to one of two groups. The rule for the behaviour of a person in the is that they will tend to relocate to a neighbourhood where there is a specific defined level of ethnicity ($f$) based on a tolerance threshold ($F$). Schelling's work determined that there is a critical threshold for $F$ where a random outcome becomes a segregated outcome.

What is interesting about Schelling's model is that it uses a simple environment of a grid of cells and a simple set of rules based on adjacent cells to generate some powerful predictions [50]. It is a good example of the use of an appropriate level of
abstraction (in terms of environment and rules of behaviour) required to answer a specific question.

One of the issues with ABM Models as they become closer to the real world is that it becomes more difficult to validate the model in terms parameters and network structure [51].

For example, a move to a continuous environment with multiple nodes representing a household would require knowledge of the connections between people in the network as well as the total number of people likely to be in a household. This is likely to be varied between different areas in a community depending on the housing and will need to be estimated using information such as census data but it may be difficult to take into account people leaving the community and entering the community.

However, being able to integrate additional information can provide further insights. For example, we may be able to see the effect of segregation where households are smaller and where different neighbourhoods within a community have improved transport links etc. There is however a growing set of principles and ideas for the processes involved in the development of ABM Models [52].

2.4.6. Development Tools

There are a number of commercial and open source software packages available for developing ABM Models. These include the following:

- Anylogic
- MASON
2. Literature Review

- RePast
- Swarm
- FLAME & FLAME GPU
- NetLogo

For choice of modelling package for ABM, language can be important as extension as customisation or development of a model beyond basic support in any of the package requires specific programming knowledge. AnyLogic is a commercial multi-paradigm tool in the sense that it supports all three popular programming paradigms of ABM, SD and DES [53]. It also allows for these to be combined with one another in a single model. AnyLogic is built on Java which is widely used and very popular. MASON [54] and RePast [55] are open source toolkits also supporting the Java language but don't have integrated support for SDM or Discrete Event Models and the environment has fewer features. Swarm [56] is open source toolkit maintained by the Swarm Development Group but requires knowledge of its own in-house programming language for developing models.

Of the popular ABM development tools, the Flame GPU [57] provides the highest performance but requires an understanding of the Flame architecture as well as knowledge of programming in the CUDA (Compute Unified Device Architecture) [58] platform developed by NVIDIA. With this it is possible to have many tens of thousands of concurrent threads running depending on the type of GPU used.

Over the years, GPU performance has gradually exceeded main CPU performance. Due to the requirement of developing around the FLAME architecture and the CUDA environment, FLAME is not an off-the-shelf product in the sense of being able to use
it to ascertain answers quickly but is an ideal long-term tool for high performance ABM. Whilst FLAME is comparatively fast, models designed using it need to be built on its architecture comprising the X-machine computational model using XML.

NetLogo [59] is a freely available application popular for building ABM and SDM models. Its main strengths are that it is easy to learn and is very flexible through extensions written using the Java programming language.

### 2.4.7. Converting System Dynamics Models to Agent Based Models

As SDM has been founded much earlier than ABM, it is often the case that we have knowledge of the SDM format of a system but need to generate an equivalent ABM Model. A number of publications have shown how to convert models from SDM to ABM. One approach by Macal [60] uses pseudo-code to describe the properties of an agent in the population as shown in Figure 3. This indicates that all agents in the system, each represented by a unique identifier, have a specific disease state (which can be Susceptible, Infectious or Recovered) and a specific recovery time. Agents in a Susceptible state have an indefinite recovery time as they are not in an infected state.

\[
\text{Agent Class} = \text{agent[agent unique identifier, agent disease state, time of recovery (if agent is in disease state “I” or “R”, if agent is in disease state “S” is } \infty]}\]

**Figure 3: A definition of an Agent in an Agent Based Model**
The state of the agent is dynamically updated. Using the agent definition, \( N \) agents are generated in states which can be one of Susceptible (\( S \)), Infectious (\( I \)) or Recovered (\( R \)). At \( t_0 \) the agents are shown in Figure 4:

\[
\text{Agents} = \{
\text{agent}[^\text{"agent 1"},^\text{"I"}],
\text{agent}[^\text{"agent 2"},^\text{"S"}],
\text{agent}[^\text{"agent 3"},^\text{"S"}],
\ldots
\text{agent}[^\text{"agent N"},^\text{"S"}]
\}
\]

**Figure 4: A definition of the states of Agents in an Agent Based Model of SIR**

During every time step, the disease state of each agent is updated by a set of logic rules shown in Figure 5 from [60].
The work by Macal and others [92][61] shows that it is possible to convert between the classic SD model and ABM using agent properties and logic. This can be done programatically or using a number of graphical based tools. One key feature of the logic is that it is very expressive, allowing it to be extended which can be useful for further experimentation.

2.4.8. Performance

One of the issues with having to repeat the simulations in ABM in order to obtain an average output for each of the time intervals is that it can quickly become time consuming. If we have a single simulation based on ABM taking 5 minutes and this

Figure 5: Update method for the ABM from 'To Agent Based Simulation from System Dynamics' [60]
needs to be repeated 1000 times in order to obtain average readings then that experiment can take over 83 hours in total. If there are a number of different experiments that need to be performed then this can become unfeasible for a project. At 83 hours and just 100 experiments, this would take almost 346 days of computation time alone.

One additional issue in the performance of ABM in terms of the time taken to complete simulations is dependant on the size of the population. One solution for this is to apply scaling to the population size. One method uses what is known as 'super-individual' populations [62]. This is where several people are grouped into one collection for the purpose of modelling. Although performance can be increased, this does however mean that the output from the ABM Model is no longer the result from local interactions of each individual within the population based on the original model. An alternative is to reduce the population size in order to make it computationally feasible. This can be performed during the investigative and initial research phases of the experiments.

In the best scenario the time required to complete a simulation based on ABM increases linearly however, depending on the interactions and the type of network connections, the time taken can lead to a combinatorial explosion making computation time unfeasible [63]. One useful method for helping to determine the most suitable level of population size uses the The Buckingham Pi Theorem [64] which takes into account the different parameters of the model in order to determine the appropriate level of scaling and this has been applied to both SDM [65] and ABM [61].
The following items provide a summary of the main reasons why ABM is inherently slow [66].

- The dependence of the SDM on random fluctuations requires many simulation runs.
- Multi-agent simulations may involve a large number of Agents.
- The simulation of rational behaviour or of human cognitive and psychological processes is potentially quite resource demanding.
- The calibration of model parameters to empirical data also requires many simulation runs for various parameter combinations to determine the parameter set(s) with the minimum error.
- The parameter space needs to be scanned to determine possible system behaviours.
- Performing scenario analyses requires many additional simulation runs.
- Visualization of simulation may be time consuming.

The cause of the speed issue when there are many agents is related to the way in which interactions occur between individuals within a population. In the real world, these interactions occur concurrently. When these real systems are transferred to computer systems, these systems are executed on microprocessors that are largely designed to run sequentially.

To work around the issue of performance, workstation resources can be increased, parallel techniques can be applied or scaling down of the population size can be attempted.
For workstation resources, the main factor is the serial nature of current processors or cores. Modern personal computer processors can have between 2 to 6 physical cores (e.g. the Intel Core i7 3970X has 6 physical cores) allowing for up to 12 virtual cores. Applications such as AnyLogic are able to process an ABM making use of each core to process the simulation concurrently. This scales to approximately 12 times the performance compared to a single core. Coupled with faster and larger memory sizes and use of a Solid State Drive instead of a conventional hard drive, this can make some simulations feasible, dropping the simulation time. Nevertheless, where population sizes are in the order of millions of persons and small world topologies are used, this can still unfeasible as each simulation can take many hours or days to complete.

The second option to increase performance of ABM is the use of parallel processing techniques. Several large scale ABM systems have been developed. One well-known available system is the Flame modelling system. Flame using a formal method using the 'X-Machine' which is an extension to finite state machines and involves automaton and states to represent agents. It is a general purpose ABM modelling system which is able to make use of a range of networked computers to carry out ABM in parallel. Therefore several hundred high performance computers can be networked to allow ABM to be scaled. This goes some way to alleviate the performance [67] but unfortunately the system is still highly sequential in nature, expensive and can be impractical due to the number of computers required for very high population sizes (such as those in the order of billions of agents).

To circumvent the issue of the limitations of networking large computers for ABM a number of ABM simulations have been run using computer chips designed for
graphics. These include EpiSimS [68] and Flame20. Flame makes use of the General Processing Unit (GPU) inside computer graphics cards to run the simulation in parallel alleviating some of the difficulties of time constraints running simulations. Flame has usability issues and the network topology that the agents run in requires writing for the simulation. EpiSimS has been used to understand spatial dynamics in a large artificial society with the population reduced to block group level.

Another highly distributed platform for high scale ABM method involves using the Global Scale Agent Model developed by Parker and Epstein [69]. This is able to use parallel execution strategies and has been shown to model infectious diseases where the population sizes is in the order of billions. The simulation in their example is developed using a population size of 6.57 billion agents on 32 nodes (one thread per node utilising 2.4Ghz AMD Opteron 8216 CPUs) and 224GB of total memory. There were 2.4 billion infections with the simulation completing in less than 8 hours. The simulation environment was based on 27,500 blocks representing 20km square regions. Each block population is scaled down for the desired total population.

AnyLogic, developed by XJ technologies allows for ABM to be built on personal computers based on a pre-developed software framework for agents. Users are able to define the properties for the agents and the agents can be placed on pre-defined network topologies such as small world or random. The application is based on the Eclipse IDE, with any specific properties or code written in Java to develop or extend the existing framework. Therefore it provides the means to run simulations out of the box with minimal effort required to set up the simulation. Focus is made on developing the properties of agents rather than focussing on the infrastructure around the simulation.
Unlike Flame GPU, the simulations are run in parallel on the main processor with the simulation also being able to use any additional cores in the computer. When the software is able to run a single simulation in parallel, each additional core has the effect of halving the simulation time but the software is not able to make use of clusters of network computers. Comparable software for ABM includes Repast and NetLogo but AnyLogic has the additional advantage of being able to build a model from different paradigms into a single simulation. Therefore equivalent models may be built using different modelling paradigms or several paradigms may be used together. Modelling very large population sizes in the order of millions or billions of agents with AnyLogic or comparable desktop software is generally impractical due to the length of simulation time which can go into days or weeks for completion. Simulations in AnyLogic must be based within the confines of the framework it provides. It cannot be extended using custom frameworks to increase performance.

The final option for overcoming the issue of performance is by making the simulation simpler so it can be completed quicker. One option for this is by scaling down the population size so that it completes within a feasible period of time using preliminary experiments. In an ideal world, the full population size is better to use. Once the population size has been scaled down it is important to validate and verify the model so it still retains the properties associated with the real world system. This adds an extra step to the overall experiment but makes the experiment feasible. In addition to the use of individual modelling, super-individual modelling may be used to increase the performance [70]. This uses groups of people rather than individuals. An alternative option is to simplify the network topology. Although the small world and scale free network topology is a popular choice for application in the SIR model,
alternative topologies can be used which are computationally simpler to run. Examples are the randomly connected network topology. This does not use local hubs for interactions simplifying the complexity of the simulations. This may meet the needs of some scenarios making ABM a feasible option. The work of Rahmandad et. al. compares ABM and SD in the context of network topologies.

2.4.9. Summary

ABM is a comparatively new technique compared to mathematical expressions. ABM can be very expressive and flexible as they use programming languages that are able to provide a wider range of expressiveness and make use of a variety of input forms. Although often used in isolation, there are software packages available to allow it to be easily combined with SDM.

As real world systems are often forms of 'complex systems', ABM is a suitable method to study such systems because:

- They have properties of self-organisation
- They are able to remain robust despite small perturbations
- They can show emergent properties which may not be apparent from their individual counterparts
- They are able to capture aspects of real systems where small perturbations can lead to phase transition [66] where the system behaviour is very different to that expected

There are several strategies for dealing with large scale simulations based on ABM such as those associated with the modelling of infectious diseases at endemic or pandemic level. One is to apply extra workstation resources, use clusters of networked
high performance machines with multi-core CPUs and large amounts of memory to tackle the issue or high end GPUs which are able to perform the simulation in parallel. Reducing the number of agents to a level that makes the simulations feasible may be an option for simplification as well as aggregating the number to 'macro' level which represent clusters of agents.

Several preliminary experiments can be carried out with the aim of identifying the largest population size that can be used in the most feasible length of time. One additional route is to simplify the system but this is not always an option. One example is by changing the topology from a small world network to a randomly connected one. In some scenarios this would not be a feasible option for studying the spread of infectious disease because the small world network is more closer related to the way infection is spread and the randomly connected network does not take into account clustering of populations. Therefore the choice of topology depends on the nature of the interactions in the system being modelled.

2.5. System Dynamics

2.5.1. Introduction

SDM is an approach to modelling which looks at how the different components of a system change over time [71]. SDM was introduced in 1955 by Jay Forrester of MIT [72]. Forrester visited General Electric and analysed the way the household appliance plants ran several shifts followed by staff layouts. Initial thoughts were that the cycles of fluctuating demand were the cause of this pattern but he went on to understand how decisions were made regarding the hiring of staff and maintenance of inventory. Using paper-drawn simulations for inventories, employees and orders, he was able to
make a prediction on the number of staff that needed to be hired for the following week and determined that if there was an oscillating condition it was internally rather than externally driven as a result of decisions made internally.

In future years, Forrester developed DYNAMO (DYNAmic MOdel) [73], a language written in a version of Algol [74], transferring hand-written calculus to computer simulation. He introduced the term dynamic to indicate that the models were not static but evolved over time. These days we can see a whole plethora of computer software dedicated to solving calculus based problems but Forrester developed these early ideas in the 1950s using his own custom software.

Forrester continued his work on SDM and in 1969 published Urban Dynamics [75] which moved from corporate systems to social systems. At the time it was widely believed that the urban issues of inward migration and loss of resources was beyond the control of cities. Forrester used SDM to explain the urban issues develop from feedback processes within the city itself rather than as a result of inward migration and helped to develop policies to mitigate some of the issues that were being identified. There was some additional work published in the mid-1970s in response to feedback on the book.

Forrester's Urban Dynamics defines the city as a complex social system where individuals in the city interact to achieve their personal goals rather than a city where the goals are not specified on individual terms but on much higher levels of scale. The idea behind this was that focus on the individual interactions could increase the overall health of the inhabitants of the city, reduce poverty and create better economic opportunities.
2.5.2. Causal Loop Diagrams

Forrester's vision embraced that of 'Systems Thinking' [76]. This is the idea of looking away from individual events and thinking about the system as a set of dependent parts, each having some kind of impact on the other,

SDM models are described using causal loop diagrams [77] which comprise directed graphs of cause and effect using nodes to represent the different elements of a system and arrows to describe the positive and negative feedback links. Delays in the system can be represented by a cross over an arrow from one node to the other.

The effect of the positive and negative feedback links is shown in Table 1.

<table>
<thead>
<tr>
<th>Type of Link</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>If the starting node increases, the destination node also increases and vice versa.</td>
</tr>
<tr>
<td>Negative</td>
<td>If the stating node increases, the destination node decreases and vice versa.</td>
</tr>
</tbody>
</table>

**Table 1: The different types of links in System Dynamics**

The positive and negative feedbacks in the nodes generate a variety of different outputs. These can be increasing or reducing, balanced, oscillating or a combination of several [78] as shown in Figure 6.
2. Literature Review

(a) Increasing Output

(b) Balanced Output

(c) Oscillating Output

Figure 6: The different profiles of SD Output [66]
2.5.3. Stock and Flow Diagrams

Stock and Flow diagrams are related to causal loop diagrams but allow for more detail through the use of a stock which represents a discrete value that is incremented or decremented over time, a flow which represents the positive or negative feedback and a delay.

An example of a Causal Loop diagram is shown in Figure 7, using AnyLogic.

![Figure 7: Causal loop diagram for births and deaths](image)

When the diagram is converted to stocks and flows, specific variables are identified and shown in Figure 8.

![Figure 8: Stock and flow diagram for births and deaths.](image)

In the simplified example of births and deaths, the causal loop diagram identifies that the population is determined by births and deaths and that there are positive and
negative influences on the births. The diagram is abstract but sufficient to capture the key components in terms of influence.

The clouds represent the input and output stocks to the model. The valve symbol indicates the flow from one stock to another. The curved arrow indicates the influence on the stock which can be based on a specific parameter such as birth rate.

The root structure of an SDM model is a set of non-linear, coupled first order differential equations of the form shown in (1) with \( x \) as the stock level, \( p \) as the parameter set and \( f \) as the non-linear function as shown in Equation 3.

\[
\frac{dx(t)}{dt} = f(x, p) \tag{3}
\]

Running the simulation involves splitting the time to discrete intervals, of \( dt \). The simulation is then stepped by \( dt \) intervals. The state of the variable is determined by its previous state and the rate of change.

\[
x(t) = x(t - dt) + dt(x'(t - dt)) \tag{4}
\]

The key element of SDM is feedback within a system. Over time, this feedback generates a result which, without running the simulation, would be very difficult to understand from the result itself. As feedback loops are the key component of SDM, causal loop diagrams and stocks and flow diagrams are essential to understand the structure of the model and for describing how it works to others. The output from SDM can be considered as the 'emergent' product of a complex system so SDM fits within the ideas of complex systems theory of complex interactions at the lower level leading to emergent products at the higher level.
Viewing the feedback diagram on its own is not sufficient to predict the behaviour of the system. Complex systems can be very dynamic and the output can change over time. Therefore simulation needs to be carried out to understand the output.

Stock and Flow diagrams are very useful for describing a model as key components and dependencies in terms of feedback are clearly visible. There are a number of documented techniques for generating stock and flow diagrams from causal loop diagrams and also from other sources such as natural language.

2.5.4. Applications

SDM is a popular choice in urban development modelling and for engineering-related modelling. A city is a complex environment with many socio-economic and environmental concerns. Special types of policies need to be considered to assist in the balancing of these needs rather than relying on market conditions or internal mechanisms to regulate the city. This was one of the early views mentioned in Forrester's work. SDM has been put into use to gain an insight into the cause and effect relationships of transportation and pollution in order to decide how best to tackle the issue of air purification in Taipei including green area preservation and public transport.

A key item identified in the research revealed that increase of public transport alone could not control air pollution and that key factors included private vehicles of employees living out of the city and in the suburbs does not balance the level of public transport on offer and that preservation of green land can be more effective that the facilitation of public transport.
SDM has been used for a variety of engineering areas including planning of electrical supply which includes a forecast of capacity, the actual delivered capacity, revenue, price and actual consumption. This has allowed engineers and researchers an insight into the effect of feedback loops in electrical supply models. One of the feedback loops this has helped to identify involves the electric rates and the consumers' reaction to the rates.

Electricity regulators review an indicated price for electricity which would allow utility companies to make revenue. After a regulatory review which involves a delay, an actual price is set. If the actual price is higher than of that proposed then there may be a reduction in the consumption of electricity after a delay where consumers review the options which could lead to an increase in the indicated price which then further declines the electricity consumption leading to a circular pattern. Identification of such patterns in engineering or business is important as it can lead to extremities of bankruptcy or high revenues.

2.5.5. Variance

Standard SDM models have no variance in the output results. This is not by limitation but by design.

In order to introduce variance in the output, a number of different strategies can be taken:

- Monte Carlo Simulation
- Stochastic differential equations
- Filters
Monte Carlo simulation involves multiple selections of random input parameters based on their distributions in order to obtain the statistics of the output variables [79]. It is useful when undertaking ‘what-if’ scenarios to undertake studies in experiments where the results are not known. In doing so, it's possible to make use of the most common case as well as best case and worst case scenarios.

One area of research identified by Hagenson [80] identifies that variance plays an important part of reducing the risk in a system using an airfield damage repair capacity in a Norwegian airbase. The risk involves mission out the uncertainty in the output space which arises due to the mechanism of feedback loops which contain uncertainty but which aren't represented in standard System Dynamic models. It is concluded that due to the feedback and randomness involved in the system, two peaks can result in the output space rather than a uniform output with a single peak. In the case of using SDM without Monte Carlo Simulation, this would result in a single peak which would not be very useful whereas with Monte Carlo simulation, this would provide a range which would be more useful.

One other type of scenario where Monte Carlo simulation is useful for SDM is where the actual input parameters are not fully understood and need to be estimated [81]. Commonly these input parameters are hand-prepared but sometimes this is automated by using computers to reduce the weighted sum of the function by the error term using a variety of parameters until the best-fit estimate is achieved. One aspect of SDM which makes this possible is the simulation speed. Due to the relative high performance of SDM it makes it possible to use Monte Carlo Simulation to estimate one or more input parameters quickly.
Monte Carlo simulation has been applied to SDM for the Susceptible Infectious Recovered model. The study by Kodaira and Passos [82] uses Monte Carlo simulation to estimate the reproduction number \( (R_0) \) which represents the estimate of new infections from one fully susceptible population based on the duration of infection. Formally, this is a product of the number of contacts per unit of time, the probability of transmission for each contact and the duration of infection. What is most notable about the reproduction number is that \( R_0 < 1 \) represents a disease free point and \( R_0 > 1 \) represents the critical threshold point for eradicating infection [83].

Alongside Monte-Carlo simulation, stochastic differential equations are also a popular method of introducing randomness into a model. One study uses Vensim [84] to look at the role of stochastic events in the dynamics of antimicrobial resistance [85]. The study examines the issues of the representation of average behaviour within SDM such that the same output is observed for any given input and starting condition. The study acknowledges that one method of obtaining uncertainty in the output is through the use of parameter modification (as in the case of the use of Monte Carlo methods for parameter selection) however it accepts that there is a possibility that an important output is missed. One possibility is that the fluctuations of values are partially random and partially uniform, driven by a combination of contact of cellular elements and external feedback acting on the population. A combination of stochastic and ordinary differential equations may therefore be suitable for this type of modelling such as the Gillespie method [86].

The last type of technique mentioned in this thesis for introducing variance is through the use of filters. Vensim contains support for a specific type of filter noted as the Kalman Filter [87] named after Rudolph Kalman who was the main developer of the
filter. This type of filter uses consecutive cycles of predictions and filtering based on noisy input data to add perturbations into the system. It is usually coupled with other techniques in diverse solutions such as engineering [88] and oceanography [89].

One method which has not currently been studied within SDM is the use of delay differential equations to introduce randomness. Although delays can be defined at the flow level, they can additionally be defined through the use of delay differential equations instead or ordinary of stochastic differential equations. These are equations where the derivatives at a specific time depend on a previous state of the system. They have been used in biological systems [90] and engineering [91] but have been used in limited places within the field of SDM.

2.5.6. Development Tools

Software packages for generating simulations based on SDM models include:

- Vensim
- PowerSim
- Stella/iThink
- AnyLogic

Vensim is a popular choice for SDM with licences available for free for non-commercial use together with commercial licences. It provides a sketching tool for Causal Loop diagrams and a further graphical environment for developing SD models. It also provides the ability to develop custom applications in a variety of languages including Visual Basic and C++. Vensim is able to automatically generate documentation for the model and generate a tree based representation of the model to identify cause and effects.
PowerSim also offers a graphical environment with some minor changes compared to Vensim. Circles are used to denote rates and diamond shapes are used to denote constants. One of the key features of PowerSim is its ability to include multiple models and link them together along with support for Causal Loop Diagrams.

Stella and iThink were introduced for the Apple Macintosh platform with Stella licenced non-commercial and iThink for commercial purposes. They provide a powerful graphical interface that allows easy generation of stock and flow diagrams and equations. iThink has some useful commercial level features such as XML based model files to allow interchange of SD models and enhanced stock types supporting queues for discrete processes.

AnyLogic is a commercial application based on a graphical user interface which allows users to easily drag and drop elements of SDM in a working window to generate a model. The key feature of AnyLogic is its ability to easily combine different types of modelling paradigms (SD, ABM, Discrete Event) into one single model. AnyLogic has an extensive set of libraries designed for commercial use such as the pedestrian library which is able to draw and animate pedestrians together with themed density maps visually mapping areas of high pedestrian movement.

2.5.7. Performance

SDM scales easily to any population size as it does not explicitly model individuals. The performance with a population size of a few thousand is not impacted when the population size is in the order of millions or billions because each time step involves solving differential equations. Therefore this makes the simulation of SDM possible.
on a very wide range of equipment possible, from hand-written methods to spreadsheets, calculators and full computer simulation software.

If the stocks and flow diagram has been prepared on paper, it is easy to transfer this to a computer simulation in a matter of minutes using the drag and drop functionality available in many software packages including Vensim.

2.5.8. Summary of System Dynamic Techniques

Real world systems such as cities are very complex, comprising many elements with different types of feedback loops. Looking at such systems from a higher level rather than low-level cause and effect can help to alleviate issues that lead to new problems that are difficult to spot at the lower level. SDM can provide recognition of the effect of these loops to identify these new problems and provide more realistic forecasting.

Often in real world systems, problems are found too late in the development process making it very difficult to resolve at a future point in time due to the infrastructure and policies that have already been put in place. Because SDM is fast, it is possible to generate many different models to make predictions and refine policies earlier in the process.

2.6. Related Studies of Variance using ABM and SD

One of the key features of ABM Models compared to SDM Models is that they are stochastic in nature due to the random nature of interaction between agents in the model, different output results are generated for each simulation output [92]. As an example, the two charts shown in Figure 9 illustrate two different outputs from an SIR simulation using a population of 500 people with 25 initially infected with an
infectivity rate of 5% and average illness duration of 15 days. The network topology is based on a random network.

![Figure 9: An example of how ABM Simulation output varies across two simulations](image)

Peak infection rate for the first figure shows 18 days whereas peak infection rate for the second output is at 20 days. Recover is also faster for the first simulation.

The benefit of having a different outputs per simulation is that this it can provide a range of results which may be more useful. In the case of the two simulation runs, we can see that peak infection values occur at two different time intervals. If a large number of such simulations are carried out then this can provide a range of values for peak infection. The results may be presented as box plots for each week, showing the minimum and maximum quartiles for each week together with the mean or median values.

In healthcare, in the case of infectious disease, this can be useful in planning scenarios where an estimate is required for peak infection as this may be the time in which the
greatest burden is placed in healthcare [93]. In social systems this can provide information regarding peak population flow through a city. And in traffic monitoring scenarios [94], it can give an indication about the range of values for the number of cars stopping at a roundabout or traffic signal.

It is possible to make the variation in ABM repeatable by seeding the random number generation to specific values prior to the simulation.

One key related work looking at variance has been discussed in a publication by Macal [60]. The work covers the conversion of an SD model to ABM using the SIR model as an example. The outputs from the three simulations are compared on 1000 simulation runs based on the following:

- Differential equations
- SDM
- ABM

The results from the simulation runs are shown in Figure 10 based on the work by Macal [60].
Figure 10: A comparison of different modelling solutions for SIR [60]

For ABM, the variation in the output curve for 1000 simulations is shown in Figure 11 in the form of plots for individual ABM simulation runs and the mean and median values.
Figure 11: A view of the variance across one thousand simulations for SIR using Agent Based Modelling [60]

Based on the experiment results it was possible to obtain counts for the peak infected individuals and peak infection times over the simulation runs.

One of the conclusions drawn from the work was that it is possible to ascertain additional information from the ABM because of the stochastic nature of the interactions of agents' interaction leading to transmission of infection. However, detailed analysis of the variation over time is not performed.

A study undertaken by Rahmandad [95] looks at the question of when to use AB and when to use differential equations using the SEIR model of contagious disease. Two aspects are studied: heterogeneity in the attributes of an agent the impact of its network topology on the interactions using a fully connected, random, small world, scale free and ring lattice network.

One of the key conclusions is that the ABM is able to represent the equation based model whilst preserving the stochastic nature of the model. One aspect very relevant to this research is that sensitivity analysis was performed for ABM to show the effect different network topologies have on the variation of the output results. For ABM, this is shown in the form of bands representing different envelopes for all the output curves.

The ABM in the work is regarded as the 'real world' and the parameters of the SD model are set to provide the best fit to the ABM.

A description for the different elements identified in the proceeding charts is shown in Table 2.
2. Literature Review

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{DE}_B )</td>
<td>Output from SD Model with mean parameters of AB model</td>
</tr>
<tr>
<td>( \text{AB}_M )</td>
<td>Mean of AB Model</td>
</tr>
<tr>
<td>( \text{DE}_F )</td>
<td>Best fit SD Model</td>
</tr>
<tr>
<td>Inner band</td>
<td>50% of ABM simulation output</td>
</tr>
<tr>
<td>Centre band</td>
<td>75% of ABM simulation output</td>
</tr>
<tr>
<td>Outer band</td>
<td>95% of ABM simulation output</td>
</tr>
</tbody>
</table>

Table 2: A Legend for the different elements shown in the output charts for the Agent Based and System Dynamics models from Rahmandad and Sterman [86].

The output for the small world network topology for the homogenous (left) and heterogeneous model (right) is shown in Figure 12.

![Figure 12: Output for the SEIR model using the small world network from Rahmandad and Sterman [86]](image)

The output for the scale free network topology for the homogenous (left) and heterogeneous model (right) is shown in Figure 13.
Figure 13: Output for the SEIR model using the scale free network from Rahmandad and Sterman [86]

The mean outputs for SD and ABM is virtual identical, generating the same output curve.

Rahmandad's work views two specific aspects of modelling of ABM, the different network topology involved in the interactions among agents ranging from random to organised and the heterogeneous characteristics of the agents. It was found that the sensitivity of the output depends on the topology with uniform, random and scale free having the lowest levels of sensitivity, a medium level for small world and high for ring lattice.

One key finding that also fits in with the work by Macal [60] is that ABM is able to provide additional insight into the range of uncertainty which comes from the stochastic nature of the model.

It is noted that the ring lattice is unrealistic as a form of transmission from one person to another but may be suitable for scenarios where the location of agents is static such as the plant population [96]. It was found that the heterogeneity of individuals has little influence of the dynamics of the interactions but it is able to identify two sub-populations: one highly connected and the other with sparse interactions so the epidemic is able to transfer faster in the first population and transfer slower in the
second type with the result being a faster peak followed by a slower decay of infection.

Discrete Event Systems have some similarities to ABM in that both forms of modelling are stochastic in nature and can generate different outputs for each simulation run. However the main difference is that in Discrete Event Systems, focus is on the process flow rather than the individual entities and their interactions.

Research looking at variance of Discrete Event Systems is limited but one study looks at the output accuracy of Discrete Event Systems and ABM using data from a womenswear department of a UK department store [97]. In the study, it was found that the Discrete Event Model didn't reflect true variability compared to ABM but this may have been due to the queuing structure used in the Discrete Event Model. It was also found that the fit of the output data to the real system was very close to that of the output from the ABM Simulation, including the variance.

2.7. Adding randomness to System Dynamics

Standard models for SDM have are deterministic. Randomness can be added to SDM so that the output varies in two ways. The first involves making changes to parameters outside of the model using methods such as Monte Carlo simulation to select suitable input parameters. The second involves making changes to the actual model itself without any changes to the input parameters. Internal changes can be brought about by the use of stochastic differential equations [98] instead of ordinary differential equations.

Comparisons between stochastic differential equations and ABM is limited. One such study [99] compares stochastic differential equations implemented using the Gillespie
algorithm with an ABM of early stage cancer. The study concludes that although it is possible to obtain models that are equivalent in terms of output data, the results are statistically different because the model based on the Gillespie algorithm is unable to retain any individual memory of past events.

One additional insight from the experiments is that the ABM Model is able to produce extra patterns of behaviour which can be considered as emergent behaviour. For example, in some simulations in the ABM Model, it was found that the tumour cells were able to be eliminated by the immune system. The study also concluded that the choice of method to use needs to take into account the availability of real world data as well as the computing resource for the ABM Model.

2.8. Conclusions

Computer simulation is a relatively new form of using computer models to understand real world systems. Within computer simulation, there are three distinct modelling paradigms, SDM which is based on analytical foundations, ABM based on agents and their interactions and DES, which are based on events.

ABM and SDM have different origins but are currently used in many related areas of industry. SDM has been around much longer than ABM with roots firmly in mathematical modelling but ABM adoption has grown steadily alongside technological advances in desktop computing and data measurement.

DES is a natural choice for real world systems containing events, entities and queues. ABM allows systems to be naturally built from their lowest elements, the agents and their interactions.
In ABM, agents and their interaction form the key component of the system and what gives rise to emergent properties is the interactions of these agents. ABM Models exhibit the following properties:

- Each agent acts independently of others
- There is no centralised form of control
- Computation is asynchronous

In SDM, the focus is on feedbacks within a closed boundary using the concept of stocks and flows. What gives rise to emergence is this feedback simulated over a period time.

The information in Table 3 compares SD and ABM based on a selection of core categories [100].

<table>
<thead>
<tr>
<th></th>
<th>SDM</th>
<th>DES</th>
<th>ABM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perspective</strong></td>
<td>System</td>
<td>System</td>
<td>Individual</td>
</tr>
<tr>
<td><strong>Realism</strong></td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td><strong>Flexibility</strong></td>
<td>Low</td>
<td>Moderate</td>
<td>Very high</td>
</tr>
<tr>
<td><strong>Time Domain</strong></td>
<td>Continuous</td>
<td>Discrete</td>
<td>Discrete</td>
</tr>
<tr>
<td><strong>Run times</strong></td>
<td>Very fast</td>
<td>Moderate</td>
<td>Slow</td>
</tr>
<tr>
<td><strong>Inputs</strong></td>
<td>Rates and Flow</td>
<td>Entities,</td>
<td>Agent characteristics and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>parameters,</td>
<td>interactions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>process flows</td>
<td></td>
</tr>
<tr>
<td><strong>Outputs</strong></td>
<td>Dynamics, steady</td>
<td>Wait times,</td>
<td>Unlimited</td>
</tr>
<tr>
<td></td>
<td>state values</td>
<td>utilisation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>throughput</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: System Dynamics compared to Agent Based Modelling from Denton [89]
One of the elements of deciding the modelling approach to use is the perspective. This can be at a high level (such as that in SDM and DES) or low level (such as that for ABM). Depending on the nature of the real world system being modelled, a combined approach can be useful in some circumstances [101].

As we live in a more connected, more dynamic and complex environment, the need for forecasting has become ever more important. At a social level, travel by road, rail and air has become easier than ever. Some of the issues we face with epidemic outbreaks are not as easy to contain or understand as they were a few decades ago where air travel was more limited and public transport less easily accessible.

As transport has evolved, the links between one person to another has become more complex and the spread of infection has become more rapid. Simple assumptions of contact rate using SDM in SIR for example are unsuitable in the model as they imply that the contact rate can be adapted or increased to take into account more efficient and accessible transport links.

A more realistic model may need to take into account the nature of the connections at a lower level and therefore be more representative of the actual dynamics underlying the change. One way we can think about the dynamics of contact between people over the years is that the organisation structure of the links between people has not changed but the 'distance' between the links has. If we view this at an ABM level, we can think of this as the view that the structure of the small world network topology in a population has not changed over the decades but what has changed is the distance between the links in the nodes. This is because we still live collectively in houses, we still have communities and live within cities but what has changed is how long it takes
for us to make contact. Modifying standard SIR in SDM to represent these dynamics is difficult as individuals are not represented on their own with their own unique set of properties and rules of interaction.

Better and more reliable models of forecasting can be useful in decision making processes during influenza outbreaks [102] to help to allocate public health resources in the most suitable areas and also to determine what kind of intervention to use to reduce deaths due to illness.

The standard SIR model in SDM is easy to implement and experiment with and can be extended using multiple compartments to introduce additional communities but the underlying principle of representing the dynamics of the population as one of a homogenous type means that it is difficult to capture the dynamics of contact.

ABM Models have been used in gradually growing numbers to understand the development of influenza over time. They are able to capture the underlying stochastic nature of the interactions based on well founded network topologies that form the basis of connections among individuals and key to their suitability and validation come from being able to ascertain information relating to these interactions. This is currently limited but with technological advances over time, our understanding of this is likely to improve dramatically. Currently, simulations based on ABM Models are comparatively slow to run but there are several methods to alleviate this scenario and it is likely that this will change in the future through the availability of more parallel processing architectures at a desktop level.

One key measure that would be very useful to ascertain is the peak height of influenza contraction and the associated timing of this. This could be very useful for planning
so that healthcare workers are able to predict surges for healthcare resource. SDM can supply this information but through the use of ABM, it should be able to supply these values together with a range of values for which peak infection can occur alongside the associated timing of these peaks.
3. Initial Experiments using the SIR Model

3.1. Introduction

This section contains the initial experiments and includes the experimental design of the ABM Model and SDM model used in this chapter and chapters four and five.

This chapter contains:

- An introduction including a description of the SIR Model
- The design and implementation of the SDM model for SIR
- The design and implementation of the ABM Model for SIR
- An experiment to examine the variance of ABM compared to SDM with Monte Carlo Simulation
- An experiment to examine the variance of ABM using the Population parameter at three different levels

Early Mathematical models for epidemiology such as those by Bernoulli in 1766 [103] were useful deterministic models that could be used to determine ‘what-if’ scenarios such as the change in life expectancy following the introduction of inoculation against smallpox. Further models followed including those for the SIR model proposed by Kermack and McKendrick which were stochastic and deterministic [104].

The SIR model is the most common foundation for more complex, custom models of infectious disease [105]. In this model it is assumed that an infected person attains permanent immunity and that the latent period is very brief and can be ignored. The population in the model is divided into the following groups:
3. Initial Experiments using the SIR Model

- Susceptible (S) - those at risk of being infected
- Infected (I) - those that are infected and can transmit the infection
- Recovered (R) - those that have recovered and have no further involvement

Initially there is no-one in the Susceptible group. As infection is transmitted between individuals, the number of those in the Infected group increases. After a period of time, the number of individuals in the Recovered group increases followed by an associated drop in those in the Infected group. The experimental data used in this thesis is obtained from the Russian Influenza epidemic in Sweden between 1889 and 1890 [106]. After the outbreak of Russian influenza in Sweden, questionnaires were sent to Swedish physicians to determine ascertain information about the influenza in their region. Answers were received by 398 physicians for over 32,600 individuals.

Although there are additional SIR data sources available, the data from Russian Influenza in Sweden is selected as it contains some additional information collated during the outbreak which is useful for selecting the parameters to use for building the SIR model and it also contains a single wave of outbreak for the region of data selected for the experiments. The information returned from the postcard included the following for each district:

- Date of first influenza case detected
- Peak of epidemic
- Percentage affected

The information returned from the questionnaire included the following for each household:
3. Initial Experiments using the SIR Model

- Number of infected persons
- Gender
- Age

Some of the important findings are used as the parameter values. This included the duration of the disease which was found to be between 2.3 and 9.4 days and affected 61% of the population. For the purpose of the experiment, a single area, Österlövsta, was chosen for analysis as this area has a profile which shows a typical raise and decline of infection population counts over a period of 15 weeks. The complete data for Sweden can also be used but this will extend the ABM simulation time. The data obtained is shown in Figure 14.

![Figure 14: Infected population values for the region of Österlövsta in the Russian influenza in Sweden based from Skog and Linde [96]](image)

The data was taken from parishes surrounding one of the regions, Österlövsta, based on original work by Linroth over a period of 15 weeks. The total population size is 52910 people and is used in the experiments. The other parameters for the models are
based on the main findings of the study with the illness duration set to 4.2 and the probability of infection set to 0.065. This produced a best fit for the selected region of Österlövsta.

3.2. System Dynamics Model of SIR

The SD model is based on the original SIR model proposed by Kermack and McKendrick [104]. The model captures the spread of a contagious disease in a closed population over time. Three coupled, ordinary differential equations are used to represent the rate of change of the three different states of the people in a given population. The equations in the model are shown with the rate of change for each of the components of SIR.

\[
\frac{dS}{dt} = -aSI \\
\frac{dI}{dt} = aSI - bI \\
\frac{dR}{dt} = bI
\]

The meaning of the parameters is shown in Table 4.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>Infection rate</td>
</tr>
<tr>
<td>(b)</td>
<td>Recovery rate</td>
</tr>
<tr>
<td>(S)</td>
<td>Susceptible population</td>
</tr>
<tr>
<td>(I)</td>
<td>Infected population</td>
</tr>
<tr>
<td>(R)</td>
<td>Recovered population</td>
</tr>
</tbody>
</table>

Table 4: A description of the parameters for the SIR model
From the equations, an SDM model is built in AnyLogic with stocks labelled as Susceptible, Infectious and Recovered and flows labelled as Infection Rate and Recovery Rate as shown in Figure 15.

![Figure 15: A Stocks and Flows diagram for the System Dynamics Model for SIR](image)

The SDM model produces the results shown in Figure 16.

![Figure 16: Results for the System Dynamics Model for SIR](image)

### 3.3. Agent Based Model of SIR

A description of the ABM for SIR is shown in this section using the Overview, Design concepts and Details (ODD) protocol [107][108] together with a feasibility study and initial results. This protocol describes the model in terms of the Overview,
Design concepts, and Detail and designed to make ABM easier to understand and reproduce if required.

3.3.1. Purpose

The purpose of the model is to understand the variation in the output curve generated by the model for the infected population and how the variation is affected by changes in the input parameters for the model.

3.3.2. Entities, state variables and scales

In the ABM model, each person is in one of the following states:

- Susceptible
- Infectious
- Recovered

The following variables are used as parameters representing different aspects of the infectious disease:

- Illness duration
- Contact rate
- Infection rate

3.3.3. Process overview and scheduling

The process is governed by the following state transitions:

- A person who is in a Susceptible state moves to an Infected state determined by the rate of infectivity.
3. Initial Experiments using the SIR Model

- A person who is in an Infected state is able to pass on an infection directly to another person determined by the contact rate. The person selected for passing on the infection is chosen at random.

- A person in an Infected state recovers after a finite amount of time determined by the illness duration.

- Once a person has reached the recovered state, they are unable to pass on the infection to another person.

The infection is passed from one agent to another randomly connected agent in the network to another at a fixed contact rate and an individual recovers from the infection using a recovery rate. A state chart is used to model the state of the agent as shown in Figure 17.

![State chart for the Agent Based Model for SIR](image)

**Figure 17: State chart for the Agent Based Model for SIR**

As per the SDM model, a population count of 52910 is used in the model based on the Russian Influenza epidemic in Sweden.

### 3.3.4. Basic principles

The SIR model is a type of compartmental model and this allows it to be represented easily in an ABM Model by representing each compartment as a state.

People, which are the agents in the system move from one compartment or state to another over time. The amount of time required to move from one state to another is determined by one of the parameters for the model.
3.3.5. Emergence

Each simulation from the ABM Model is able to produce different values for the number of people in each state of the model. This is due to a combination of the random selection of the next person in the population combined with the selection of a random individual at the start of the simulation. The rise and fall of infections characteristic of the SIR model can be considered an emergent property arising from the result of local level interactions of people in the simulation.

3.3.6. Interaction

Individuals (agents) are connected using a small world network topology. This is chosen as it represented a suitable route of transmission of the infection with many close connections in the network coupled with distant connections. The distant connections may be perceived as transportation links such as those by rail or sea. The small world network has been used in epidemiology [109] with a number of studies using it as part of the models [110].

3.3.7. Stochasticity

Seed values used during the experiments are fixed so that it is possible to reproduce the results. The seed values are used by AnyLogic to select the random person within the population to pass the infection to.

3.3.8. Collectives

There is a single collection used in the model which represents the population.
3.3.9. Initialisation

Initially, a single agent (person) is chosen at random for the initial infection. Subsequently, further infections are passed from this selected person to others.

3.3.10. Input

The inputs to the model are the values of infection rate, illness duration and contact rate. These are fixed values set as input parameters as part of the configuration of the model in AnyLogic.

3.3.11. Feasibility study and initial results

A feasibility study is carried out for the modelling software and AnyLogic by XJ Technologies chosen as a suitable choice for modelling SIR in SDM and ABM. One of the features of AnyLogic is that it has inherent support for combining different modelling paradigms into a single model.

In total, 100 experiments are carried out in AnyLogic. The experiments are carried out on a PC running Windows 7 with 3GB memory and an Intel Core 2 P8700 microprocessor. Output from the experiments is imported into MatLab to generate the box plots.

The result for the output of the ABM simulation is shown in Figure 18. In this figure, the statistics are shown as box plots for each week. Each box plot represents the data from 100 simulations showing the minimum value, the maximum value, the Interquartile Range (IQR) and the median value.
Figure 18: Results of the Agent Based Model for SIR

As per the SD model, the AB model is also validated against the data from the Influenza epidemic. The ABM simulation takes a total of 13 hours to complete.
3.4. Variance of ABM compared to SD using the SIR Model

3.4.1. Introduction

Classical deterministic simulations of epidemiological processes, such as those based on SDM, produce a single result based on a fixed set of input parameters with no variance between simulations. This type of output may be suitable for certain types of questions but is limited as it does not take the natural random elements of the system into account and therefore the output does not express the larger range of outcomes which may be more useful in understanding the range of peak infection quantities and the timing of those quantities.

For SDM, one way to introduce this variation in the output space is by modifying the input parameters on the simulations using Monte-Carlo methods, to understand how changes in the input parameters affect the spread of results for the simulation, as described in Chapter 2.

Simulations based on ABM are able to produce different output results on each run based on knowledge of the local interactions of the underlying agents and without making any changes to the input parameters. In this thesis we compare the influence and effect of variation within these two distinct simulation paradigms and show that the ABM simulation of the epidemiological SIR (Susceptible, Infectious, and Recovered) model is more effective at capturing the natural variation within SIR compared to an equivalent model using SDM with Monte-Carlo simulation. To demonstrate this effect, the SIR model is implemented using both SDM (with Monte-Carlo simulation) and ABM based on previously published empirical data.
Models of infectious diseases can be useful for understanding the spread of infection of the diseases within a population over time. However, within a given population, diseases can spread at different rates over time due to the natural random nature of contact between individuals in the population. If a simulation can incorporate this kind variation, the extra information can be used to determine the spread of uptake of infection in worst case and best case scenarios for a given population.

Currently, for classical SDM models based on ordinary differential equations, the random contact between individuals is aggregated to fixed rates of contact and the output has no variation. Assuming the same parameter values are supplied to the SDM simulation, on each run, the same results are produced. Subsequently, in order to understand the spread of output values, the simulations are repeated with different input parameters by applying Monte Carlo simulation [111]. In this approach, multiple experiments are performed and the parameter values taken from a probability density function representing the input parameter range. In ABM, uncertainty or variance can be inherent within the model so that the simulations from the models produce non-deterministic results directly without input parameter variation.

In this study, the two approaches are examined by generating an SD model with Monte-Carlo and an ABM and comparing the spread of output values against published data for a defined population. Simulations from modelling paradigms such as ABM, which can include variance, help to bridge the gap between raw data and simulation data and also help answer the question of validation in simulation - assessing the degree to which a model is an accurate representation of the real world [112]. Both SDM and ABM are able to capture overall variance but unlike
3. Initial Experiments using the SIR Model

Simulations from SD models, a single simulation run from an ABM Model is able to capture the ‘typical’ outcome from a single simulation experiment.

Unlike SDM which uses a top-down approach to model the system as a whole, in ABM simulations, the system is ‘brought about’ by carrying out the lower level interactions between the agents. For this reason, ABM is beginning to be used in a range of fields including biological simulations and social sciences representing people as interacting agents in environments [113].

3.4.2. System Dynamics Model using Monte Carlo Simulation

The Monte Carlo simulations are used to determine how infected population counts change when the input parameters to the SD model are varied. Monte Carlo simulation uses repetitions of random sampling of the input parameters to determine the result. The randomness is applied ‘outside’ of the internal workings of the system as it is the parameters to the system being sampled.

One of the limitations of using the Monte Carlo method applied to simulations is the time taken to perform the simulation over a very large number of iterations. Therefore in areas such as Probability Sensitivity Analysis, the Monte Carlo solution is not always a viable method for complex models such as those for healthcare, involving thousands of patients [114].

Monte Carlo simulations are carried out using the SD model to see the effect of varying each parameter and the effect of varying all parameters. In total, 100 simulations are carried out for each experiment to match the ABM. Parameter variation is carried out by randomly selecting values for each of the parameters taken from a standard normal distribution based on the mean value.
The following Monte Carlo simulations are carried out:

- Illness Duration variation
- Contact rate variation
- Infection rate variation
- Illness duration, contact and infection rate variation

Each SD experiment, comprising 100 simulations, takes a total of 9 seconds. The box plot for the SD Monte Carlo model with illness duration variation is shown in Figure 19. The infected population peaks at 21,442 people.

![Figure 19: System Dynamics Model for SIR with Illness Duration Variation](image)

In the case where the contact rate is varied, the result is shown in Figure 20. In this case, the inter-quartile range is larger than the simulation where the illness rate is varied and clearly visible in weeks 3 to 7 inclusive.
3. Initial Experiments using the SIR Model

**Figure 20: System Dynamics Model for SIR with Contact Rate Variation**

The SD Monte Carlo model where the infection rate is varied is shown in the box plot in Figure 21.

**Figure 21: System Dynamics Model for SIR with Infection Rate Variation**

The box plot for the experiment where multiple parameters are varied is shown in Figure 22. The results show that with multiple parameters being varied, the infected
population peaks at 24,725 which is a substantial increase compared with the SD version without Monte Carlo simulation which peaks at 13,025.

Therefore, compared with experiments where variations of contact rate and infection rate are altered to introduce randomness, the variation of multiple parameters has the undesired effect of scaling up the infected population counts.

![System Dynamics Model - Infection, Illness and Contact Variation](image)

**Figure 22: System Dynamics Model for SIR with Infection, Illness and Contact Rate Variation**

### 3.4.3. Validation against Influenza data

The Wilcoxon rank sum test [115] is used to compare the simulation results against the Influenza data. This is a non-parametric paired test that tests the null hypothesis that the means from the two data sets are the same versus the means from the two data sets differ.

The SD result without any Monte Carlo simulation is compared directly against the Influenza data. For ABM and SD with Monte Carlo, the median values for each experiment are obtained for each week.
The Wilcoxon rank sum test for the experiment is calculated using MatLab version R2010b. The results are summarized in Table 5. A 5% significance level is used.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>0.3013</td>
</tr>
<tr>
<td>ABM</td>
<td>0.4648</td>
</tr>
<tr>
<td>SD – Vary illness duration</td>
<td>0.2661</td>
</tr>
<tr>
<td>SD – Vary contact rate</td>
<td>0.2036</td>
</tr>
<tr>
<td>SD – Vary infection rate</td>
<td>0.0244</td>
</tr>
<tr>
<td>SD – Vary illness, contact, infection rate</td>
<td>0.0269</td>
</tr>
</tbody>
</table>

**Table 5: Wilcoxon Rank Sum Test for the SIR variance experiments**

The h value for the tests is 0 and the p values of 0.0244 and 0.0269 indicate that the null hypothesis can be rejected for the experiment where infection rate is varied and for the version in which combined parameters are varied.

The Wilcoxon rank sum tests show that the SD without Monte Carlo and the ABM has equivalent overall fits with the experimental data. The ABM experiment, with natural variation between different simulations, due to the contacts between the agents, is in agreement with the Influenza data.

When the Monte Carlo simulation is applied to the SD model, the overall results of the simulation are in agreement with the Influenza data for illness duration variation and contact variation but for variations of infection rate and the combined variation the results are no longer in agreement. The last, combined Monte Carlo simulation, has the overall effect of scaling up the median values overall.
3.4.4. Variance in ABM and SD Experiments

Variance for each of the Monte Carlo experiments are taken from the box plots and compared against the variance of the ABM experiment. In order to compare the variances, the inter-quartile range (IQR) is calculated using the MatLab for the ABM experiment and SD Monte Carlo experiments.

The IQR for the ABM is shown in Figure 23. The ABM experiment produces a broadly symmetric result reflecting variations of the uptake of the infection which occurs at different times in the simulations but producing the same shape of the infection curve.

![Figure 23: Interquartile range for the Agent Based Model for SIR](image)

There are two distinct peak values for the infected population shown in Figure 23, in week 3 and week 5. This is because there is less variation observed around the centre of the normal distribution for the results with most of the variation occurring just before or just after centre of the distribution.
The IQR for the SD Monte Carlo simulation with variations in the illness duration is shown in Figure 24. The chart shows that there is less variation at the height or peak of infections. This is because the variation is created by the random connections that the individuals have in the simulations and critical parameters such as infection rate and illness are kept constant.

The results show that ABM is able to maintain stable peak infection values whilst at the same time exhibiting the type of randomness one may expect between different populations.

![Figure 24: Interquartile range for the System Dynamics Model for SIR with Illness Duration Variation](image)

The chart in Figure 25 shows the IQR where the Contact Rate is varied in the SD Monte Carlo simulation.

The chart shows that the counts at the height of infection vary significantly between simulations compared to the ABM. In the ABM experiment, the contact rate is
constant among the simulations and therefore in those simulations there is less difference of the counts at the height of infection.

The chart in Figure 25 also has the typical shape for the SIR infection values for the population, where there is an initial sharp increase due to the initial spread of infection followed by a slow decline in the total number of infections due to people in the population recovering and no longer passing the infection.

Figure 25: Interquartile range for the System Dynamics Model for SIR with Contact Rate Variation

The chart for the IQR for the infection rate variation for the SD Monte Carlo simulation is shown in Figure 26.
3. Initial Experiments using the SIR Model

Figure 26 Interquartile range for the System Dynamics Model with Infection Rate Variation

The chart for the IQR in the case where multiple parameters are varied in the SD Monte Carlo simulation is shown in Figure 27.

Figure 27: Interquartile range for the System Dynamics Model with Infection, Illness and Contact Rate Variation
Unlike the Monte Carlo simulations where a single parameter is varied, in this case, with multiple parameter variations, there is an overall significant increase in the variation. Table 6 shows the total variation (the sum of IQR values) for the different experiments.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Total Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABM</td>
<td>16050</td>
</tr>
<tr>
<td>SD – Vary illness duration</td>
<td>10223</td>
</tr>
<tr>
<td>SD – Vary contact rate</td>
<td>19956</td>
</tr>
<tr>
<td>SD – Vary infection rate</td>
<td>26452</td>
</tr>
<tr>
<td>SD – Vary illness, contact, infection rate</td>
<td>36569</td>
</tr>
</tbody>
</table>

**Table 6: Total Variation observed for the SIR Experiments**

The least total variation for the simulation is obtained for the Monte Carlo experiment where the illness duration is varied. The Monte Carlo experiment where combined parameters are varied has more than twice the total variation of the ABM experiment.

Results from the ABM simulation showed that the overall peak total infections for the population remained stable between simulations. The shape of the output curve for each simulation has a closer fit with the curve for the empirical data. The ABM simulations only differed with the initial delay before the uptake of infection which may also arise due to natural variation of the contact rate between individuals and their connections which arise be due to the movement of individuals within the population.

In contrast, for the SD model, the effect on the variation of the parameters has the effect of altering the rate at which the infection spread within the population.
3.4.5. Discussion and Conclusion

In this experiment, the variance generated by an ABM Model with repeated simulation runs and no parameter changes is compared to an SDM Model where the variation is generated by using Monte Carlo Simulation to select the input parameters for the model.

Although variations of SD models exist which are able to integrate random elements [116] [117] they produce a different kind of variation compared to ABM. Whereas in stochastic models there is a random element applied to the equations, in ABM the randomness is inherent and more natural, following the rules of the underlying system being modelled.

The ABM and SD experiments for the SIR data show that ABM is able to capture natural variation without recourse to modification of any parameters for a simulation. The classic SD model has no variation. The SD with Monte Carlo simulation has variation but it is very sensitive to parameter changes and in the case where multiple parameters are varied, it produces variation and infected population counts which no longer match up against the experimental data. Therefore an ABM of SIR with built-in randomness is able to capture the natural variation in SIR better than a classic SD model with Monte Carlo simulation. The source of variation for the ABM is the contact between the agents between the different experiments.

Additional regions may have been used as the baseline empirical data to validate the ABM Model but this would have increased the computation time to complete all the experiments. Therefore the results for a single region as presented in order to allow a researcher to gain insight into the variation generated by the SIR model.
Several comparative studies between ABM and SD have been undertaken [118]. Some notable discussions in these studies include the issue of computing power and control. In this study, it is the case that the ABM is computationally expensive compared to the classical mathematical model although this may be overcome in future by highly parallel computing architectures [119]. Traditional continuous models are generally easier to implement but many aspects of biological systems are intrinsically stochastic in nature [120] so the ABM could be viewed as a more ‘faithful’ interpretation of the processes being modelled.

As the ABM is built using autonomous individuals, it could be extended to include connections between individuals across different regions to understand the effect of disrupting the spread of the epidemic by shutting down major transport links for example. Further work could include the effect of the use of different network topologies.

The use of ABM with its inherent and intuitive representation of natural variation and interaction among components can help to bridge the gap between computer simulation and biological systems and provide insight of how local level interactions bring about global system outcomes.

3.5. Variance of ABM using the Population Parameter for the SIR Model

3.5.1. Introduction

One of the issues highlighted with ABM is that the simulations based on ABM are much slower to run compared to equation based modelling. Therefore to make ABM a feasible option where large numbers of agents exist (such as population sizes in
regions or countries), the simulation must overcome the issue of performance. The main routes for this are to:

- Increase the amount of available resources on the workstation (e.g. increase in memory, faster CPU and SSD for storage rather than traditional hard disk drives)
- Introduce a highly parallel design (e.g. using distributed workstations, multiple processors or GPUs)
- Simplify the model - e.g. scale down the population size to make the simulation perform in a feasible length of time

The effect of increasing resources is subjective because the increase will only take effect if the package used to run the simulations is able to take advantage of the additional resource. To make the best use of this, monitoring of computer resources combined with known features of the simulation package can be used for optimal effect. In practice this is somewhat limited because desktop computing resources are finite and often the largest bottleneck is the CPU.

This was traditionally single core but current quad-core CPUs have become more commonplace for desktop computers. The introduction of additional CPU cores has the effect of reducing the simulation time by a factor of two to four but depending on the total number of experiments, this improvement may not be sufficient.

Several large scale ABM systems have been developed using highly parallel designs. One well-known available system is the Flame modelling system. Flame uses a formal method using the 'X-Machine' which is an extension to finite state machines and involves automaton and states to represent agents.
It is a general purpose ABM modelling system which is able to make use of a range of networked computers to carry out ABM in parallel. Therefore several hundred high performance computers can be networked to allow ABM to be scaled. This goes some way to alleviate the performance but unfortunately the system is still highly sequential in nature, expensive and can be impractical due to the number of computers required for very high population sizes (such as those in the order of billions of agents).

ABMs based on highly parallel systems may therefore offer a potential route for large scale simulation. However all of the systems that currently employ highly parallel designs use specific frameworks and rely on specific types of hardware. These are not typical 'off the shelf' products. Therefore they may not be a feasible option where there are time, cost and knowledge constraints.

The final approach mentioned in this thesis is to simplify the simulation by scaling down the number of agents or the population size (where the agents represent people in the simulation). This can be done using standard 'off the shelf' desktop simulation software without the need for any additional knowledge or hardware.

The scaling factor can be determined by sensitivity analysis and by the needs of the experiment. In an ideal experiment, the full number of agents is the best number to use as it ties closest with the real world. However, computer simulations are abstractions or simplifications of the real world to make it possible to answer specific questions therefore a simplification of the population size may be one of the abstractions to make it possible to carry out a study.
Ideally, the population size should be as close to the real-world scenario as possible. This is because the effect of random variation on the output space is higher in smaller populations than it is for larger populations.

### 3.5.2. Software and Hardware Used for the Experiment

AnyLogic is used to generate the ABM model. The experiments are carried out using a Windows laptop computer with the specifications shown in Table 7:

<table>
<thead>
<tr>
<th>Item</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating System</td>
<td>Windows professional 8.1 64 bit</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel Core i5-2540M 2.6GHz 2 core with 4 virtual cores</td>
</tr>
<tr>
<td>Memory</td>
<td>12GB</td>
</tr>
<tr>
<td>Drive</td>
<td>Intel SSDSA2BW160G3H 160GB SSD</td>
</tr>
</tbody>
</table>

Table 7: Specification of the computer used to perform the scaling experiments

### 3.5.3. Agent Based Model used for the Experiment

The initial experimental set-up involves generating the ABM and validating it against the empirical data using a Wilcoxon rank sum test. For the small world network, 9 connections per agent are used with a neighbour link fraction of 0.94. These have been generated by repeated simulations with different parameters across a range to determine the best fit.

In total, 100 different simulations are carried out and the median values for the simulations are used to apply the validation. This is because for ABM, each simulation generates a different output curve due to the random interactions that occur
between persons in the population during the simulation. The output curve from the initial model provides a baseline for the further experiments carried out.

Three sets of experiments are performed to examine the effect of population size reduction:

- Reduced population size only
- Reduced population size with environment change
- Reduced population size with environment change and simulation delay

In each of the three sets of experiments, two different population sizes are used. The first uses the population scaled down by a factor of 10 representing a population size of 5291 and the second uses the population scaled down by a factor of 100 representing a population size of 529. Therefore there are a total of 6 experiments split into 3 sets.

The first set of experiments does not apply any changes to the model, only the population size is changed. The first set of experiments is used to determine the effect of reducing population sizes on the output curves.

In the second set of experiments, the environment is changed so that the output curves from the experiment are closer to the expected outputs for the specific population size being used. The connection per agent is increased to 30 and the neighbourhood link fraction set to 1.0.

In the last set of experiments, there is a new environment with connections per agent set to 50 and a neighbour link fraction set to 0.94. It is combined with a time delay so that the time when the peak infection occurs matches that of the empirical data.
Because the output curve of each simulation in the experiment is different, 100 simulations are performed for each experiment and the output from each of the three experiments is shown in a box plot. The median value is also shown for each experiment. The box plots are used to compare the output space and statistics of the three different experiments.

3.5.4. Initial Model

The infection is passed from one agent to another randomly connected agent in the network to another at a fixed contact rate and an individual recovers from the infection using a recovery rate.

The results for 100 simulations are shown in the box plot in Figure 28. This shows the inter-quartile range for each week across the 100 simulation runs as well as the median value for each week.
3. Initial Experiments using the SIR Model

**Figure 28: Output from the Initial Agent Based Model for SIR**

The results show a sharp rise in the total number of people infected between weeks 3 to 5 followed by a gradual reduction.

In this model, which is a basic implementation of SIR, everyone eventually recovers from the infection. In real world systems this is not always the case as it is possible in some circumstances to never fully recover from infection, therefore in these circumstances, the profile of the curve would be wider and peak number of infections would not drop to a value of 0. However this model is an abstraction of the real world SIR for the purpose of undertaking research in variation.

The empirical data for the Russian Influenza in Sweden is shown in Figure 29. The output curves from the initial model shown in Figure 28 are comparable to the empirical data but rather than having a single value for the infected population for any specific week, there are a range of values which are represented as a box plot for each week.

The comparison of the output from the ABM Model to the empirical data is important for the purpose of model validation, to validate that the output of the simulation is consistent with the output from a real world system. There are currently a number of studies to investigate validation in the field of ABM [121]. In this research, a Wilcoxon rank sum test is used as this allows the output values of the simulations to tested directly against the empirical data. This form of testing has been used in a number of studies to validate and compare data resulting from different types of models [122].
3. Initial Experiments using the SIR Model

3.5.5. Results for Reduced Population Size with no Environment Change

The model is reduced by a factor of 10 from the initial setup bringing the population size to 5291 and the experiment is repeated. The results are shown in Figure 30.

The results show that the peak infection occurs much later, in week 11 compared to the empirical data. This is the result of reducing the population size by a factor of 10 but keeping the environment settings consistent with the original experiment.

As the people in the population are now more spread within the environment at the smaller scale, it takes longer for the infection to spread within the population; therefore there is a slow rise to peak infections compared to the empirical data where there is a rapid rise. Recovery of the total population does finally occur but this occurs much later in time in the experiment.
Figure 30: Output from the initial Agent Based Model with Population size 5291

The final experiment in the first set of experiments uses a population size of 529. The results are shown in Figure 31.

For the population size of 529 people, the peak infection occurs between weeks 6 and 9. Compared to the peak infection for the population size of 5291, the spread of peak infectivity occurs within a larger number of weeks.

Compared to empirical data, there is a linear increase in the simulation output to peak infections whereas in the empirical data there is a rapid increase. There is also a rapid decline in infections compared to empirical data where the decline is gradual.
3. Initial Experiments using the SIR Model

Figure 31: Output from the Initial model with population size 529

Peak infectivity occurs earlier compared to the output for 5291 people, occurring between weeks 6 and 9 instead of week 11. This may be the result of a smaller population size. So although the environment is the same for both the 5291 and 529 population size, infection may be spreading faster where the population size is 529 due to the small sample size.

3.5.6. Results for Reduced Population Size with Environment

Changes

The links per agent in the environment is increased. The results for population size of 5291 combined with a change in the environment to increase the value for the infected population are shown in Figure 32.
Figure 32: Output from the Agent Based Model using updated environment with population size 5291

With the increased connections, the peak infection occurs much earlier, in week 4 instead of week 11 shown in Figure 30. This makes the output of the simulations for population size 5291 valid against the empirical data.

The results for population size of 529 combined with the change in the environment to increase the population size are shown in Figure 33.
3. Initial Experiments using the SIR Model

Figure 33: Output from the Agent Based Model using updated environment with population size 529

With the environment change for a population size of 529 people, the peak infection occurs in week 3 compared to the version without the environment change where peak infection occurs between weeks 6 and 9. This makes the output of the simulations for population size 529 valid against the empirical data. Also, in the version without the environment change, there isn't a single peak infection but rather it is spread across several weeks. With the environment change, peak infection occurs during a single week in line with the empirical data.

3.5.7. Results for Population, Environment and Time Delay Changes

In the results with the environment change, peak infection occurs a little earlier than expected compared to the empirical data. This is due to the change in the number of connections for each person in the population which results in larger clusters of
3. Initial Experiments using the SIR Model

groups of people which results in a rapid increase in the number of infections. To resolve this issue, a time delay is used to push forward the week in which peak infection occurs. This means that the curve for the output from the simulations lines up with the curve for the empirical data.

The results for population size of 5291 combined with both the environment change and the time delay are shown in Figure 34.

![Environment change with Warm up - Population size 5291](image)

**Figure 34: Output from the updated environment using delay with population size 5291**

With the time delay, peak infection occurs in week 5 in line with the empirical data. Also, the shape for the output is much closer to the empirical data.

The results for population size of 529 combined with both the environment change and the time delay are shown in Figure 35.
3. Initial Experiments using the SIR Model

3.5.8. Analysis of Results for Stepping of Population Parameter

In the first set of experiments, the peak value for the total infected population was far below those expected at one tenth scale. The actual peak infection for the empirical data lies between 12000 to 14000. At a tenth scale we would expect this to equate to between 1200 to 1400 whereas the actual output for the peak value is in the range 300 to 400.

Similarly at one hundredth scale we would expect the peak total infected population value to lie in the region of 120 to 140. The actual peak total infection for the experiment is in the region of 50 to 60.

Figure 35: Output from the updated environment using delay with population size 529
3. Initial Experiments using the SIR Model

In the first experiment, peak infection value is achieved in week 11 compared to week 5 for the empirical data. Therefore in addition to the small peak infection value, the rate of infection is also much smaller in comparison.

At one hundredth scale, peak infection was achieved at 8 weeks which is closer to the empirical data in terms of the rate of infection.

For the first set of experiments, we can conclude that the output curves do not match the empirical data.

In the second set of experiments, the environment is modified to increase the rate of infection to take into account the smaller population size. In this instance, the peak value for one tenth scale is close to 1400 which is aligned with the expected total infected population value at this scale. At one hundredth scale, the total infected population value is 140 which is also aligned with the expected output at this scale. There are 8 to 10 weeks between the start of infection and recovery.

Therefore, for the second set of experiments, the output curves closely match those of the empirical data with one exception: the number of weeks before peak infection occurs.

The final set of experiments take the point of peak infection into account by introducing a time delay. This delay means that the peak infection value occurs later in time.

For the final experiments, the environment modified in combination with a delay. At one tenth scale, peak infected population is 1400 persons but this time, peak infection occurs during week 5 which is the expected time for peak infection, aligned to the
empirical data. At one hundredth scale, peak infection is between 160 to 170 and also occurs at week 5.

For the final experiment, there are 8 to 10 weeks between the start of infection and recovery which matches the empirical data.

For the final experiments, the simulation is running at one hundredth scale generating output curves that match the empirical data. We find that a time delay combined with an environment change is necessary as the population size is reduced.

3.6. Conclusions

In this experiment, the ABM Model is implemented in SIR and the population parameter is adjusted at three orders of magnitude. It is found that it is possible to reduce the population size whilst maintaining the relative level of variation among the different values for the population size.

Therefore, in the event that processing time is an issue at full scale population sizes then it may be possible in the case of the SIR model, to use smaller population sizes combined with environmental changes and a time delay to make experiments run within a feasible amount of time. This could make experiments involving entire countries or regions feasible at a much smaller scale making it possible to run experiments using standard desktop computers with minimal training and hardware requirements. Such smaller scale models may make help scenarios where the different strategies of vaccination for influenza need to be assessed on a cost-effectiveness basis.

It is also found that the main environmental change required is an increase in the number of connections per agent as the population is decreased.
Although it was seen that the output curves matched the empirical data at one tenth and one hundredth scale, further research needs to be carried out to determine:

- If the range of output values for each week (representing the variance between the total infected population) match those across the different population sizes.
- If the properties of the output curves are maintained in models where there are a series of peaks and troughs.
- How the output curves change with different populations with other network topologies (e.g. random connections). Similar research has been carried out with ABM and SD.

This is investigated further in Chapter 4.
4. Variation of ABM and SD across Stepped Parameter Ranges

4.1. Introduction

In the initial experiment, the variation in the output space of an ABM Model is compared to the variation in the output space of the SD. The input parameters are not modified in the ABM Model but are modified in the SDM model using Monte Carlo simulation. This allows the natural variation which is internal to the ABM Model to be compared to the variation in the SDM model applied outside of the model itself using the input parameters.

This experiment examines how the variation in the output space varies in SDM and ABM when a defined range of input parameters are covered. Rather than randomly selecting the input parameters using Monte Carlo simulation, the input parameters are incremented for both models. This way there is consistency among the parameter values and a fairer comparison can be made for the two models (i.e. the same input values are chosen for both sets of SD and ABM experiments).

For each parameter value that is incremented, a Wilcoxon rank sum test is performed using the output data from the simulation against the empirical data. Initially the test results indicate that the two sets of data are a match. As the parameter is incremented, the test results move from a state of match to a state of non-match. This is referred to in this research as the tipping point within the context of data matches and non-matches. The tipping point is the point at which the output curve from the simulation no longer matches the output curve from the empirical data for the selected parameter.
4.2. Variation of ABM and SD for SIR

4.2.1. Introduction

For the purpose of the experiment, a single area, Österlövsta, was chosen for analysis. Initial research indicated that it would not be feasible to run the simulations for all areas as the population sizes due to the length of time taken to complete the simulation. This Österlövsta area has a profile which shows a typical raise and decline of infection population counts over a period of 15 weeks.

An overview of the process for the experiments is shown in Figure 36:

![Figure 36: Overview of the process for the tipping point experiments](image)

Two sets of experiments are performed for each model and for each parameter. In the first, the parameter is decremented until it reaches its lower value. In the second, the
4. Variation of ABM and SD across Stepped Parameter Ranges

Parameter is decremented until it reaches its upper value. In each iteration of the parameter variation, validation is performed against the real data. At the end of the experiments, the results are analysed using box plots.

The illness duration parameter is incremented from 1.0 to 10.0 in increments of 0.1 to determine the effect of parameter variation on the output space for SD and ABM. The process is repeated for contact rate from 0.1 to 20.0 in increments of 0.2.

A special parameter, the total population is also selected for variation. The experiments are repeated for population counts of 5291 and 529. These experiments show some of the implications of scaling SD and ABM at different population counts. This is important because ABM computation time is not always feasible for very large numbers of agents. The impact of scaling up the population size is discussed in the next subsection of this thesis (4.2.2).

An SD and ABM is generated and the output curves from the simulations are validated to the Österlövsta influenza data by applying a Wilcoxon rank sum test with the output from the models against the empirical data. The test is of a nonparametric nature and can be used to determine whether the distribution of data between two separate outputs on a dependent variable is different from one another.

A feasibility study is carried out for the modelling software and AnyLogic by XJ Technologies chosen as a suitable choice for modelling SIR in SDM and ABM. AnyLogic 6.9 is used for the experiments. AnyLogic is easy to use due to the way that a model can be developed using the user interface and with less coding required to complete simple models but the models can be extended using the Java programming
language. As it runs on standard desktop computers, models can be developed using existing technology.

One of the additional features of AnyLogic is that it has inherent support for combining different modelling paradigms into a single model if required for further research.

As per the SDM model, a population count of 52910 is used in the model based on the Österlövsta area. Individuals (agents) are connected using a small world network topology. This is chosen as it represents a suitable route of transmission of the infection with many close connections in the network coupled with distant connections. The distant connections may be perceived as transportation links such as those by rail or sea. The small world network has been used in epidemiology with a number of studies using it as part of the models. A single randomly connected agent is chosen to kick-start the spread of infection.

For the full population of 52910 individuals, no warm-up time is required for the simulation. The warm-up time is the time required for the simulation to progress before the results can be gathered. For the population values of 5291 and 529, a warm-up time is introduced. For the smaller population sizes, the spread of infection to its peak infectivity occurs too quickly compared to the full population. Therefore the warm-up time allows the curves for the different population sizes to match up.

The result for the ABM with the full population is shown in Figure 37: Output for the Agent Based Model for SIR with a Population of 52910. Each single simulation with the full population set took approximately 4.2 seconds to complete.
4. Variation of ABM and SD across Stepped Parameter Ranges

Figure 37: Output for the Agent Based Model for SIR with a Population of 52910

The experiments are carried out with the full population set of 52910 people and then repeated at one tenth scale and one hundredth scale using values of 5291 and 529 respectively. The illness duration value is stepped from 1.0 to 10.0 in increments of 0.1 for both SD and ABM at full population level. This is to obtain data for the amount of time taken and the tipping point results at full scale.

Smaller increments would be more ideal as that would allow for a more thorough analysis of the results. This is not practical due to the computation time required to carry out the experiments but it is possible to carry out a more detailed hypothesis test by 'zooming' into the tipping point region. This can be used to identify more detailed tipping point values. The process is repeated for contact rate where the value is stepped from 0.2 to 20.0 in increments of 0.2. Only a single simulation run is required for SD because the same output curve is generated for a specific set of parameters.

For ABM, for each increment for the illness duration value, the simulation for each stepping value is repeated 100 times and an average median value obtained. This is to take into account the variation in the output results for ABM.
A Wilcoxon rank sum hypothesis test is carried out comparing the output value of the simulation for each of the stepping values against the original empirical data. The results are shown in the form of a banded chart with the dark bands represented $h=0$ and the light area representing $h=1$.

The ABM hypothesis tests are carried out using the same stepping values and the output from the simulations.

### 4.2.2. The impact of scaling up the population size

In order to understand the impact of scaling up by a factor of 10 rather than scaling down by a factor of 10, a number of initial simulations were carried out using a total population size of 529,100. Animations and charting during these experiments is turned off to reduce any delays caused by user interface updates. The time taken to complete a single simulation for a population size of 529,100 took 14,314 seconds which equates to 3.976 hours. Using this figure, the total number of experiments required are shown in Table 8. The total simulations column represent the number of stepping simulations carried out combined with the number of repeated simulations carried out for each stepping value.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Total simulations</th>
<th>Time required (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illness duration</td>
<td>91,000 (91×100)</td>
<td>36,182 (9,100×3.976)</td>
</tr>
<tr>
<td>Contact rate</td>
<td>10,000 (100×100)</td>
<td>39,760 (10,000×3.976)</td>
</tr>
</tbody>
</table>

Table 8: Time required for running stepping experiments for a population size of 529,100

Based on this, the total time required to complete all experiments is 75,942 hours or 3164 days. This does not include any further experiments required to make any
4. Variation of ABM and SD across Stepped Parameter Ranges

adjustments which may be required for the environment in order to validate the output curve from the simulations against empirical data. There are a number of factors which cause this delay including the choice of network topology used (small world network) and the software application used to carry out the experiments (AnyLogic) which is more efficient at carrying out smaller scale ABM simulations.

4.2.3. Experiment for Illness Duration Parameter

The results are presented here for illness duration with sections for each of the three population sizes. The results for each type of SD and ABM simulation, a box plot is shown with the values for each of the 15 weeks of the simulation. These range of values represent the values that exist for each week as the parameter is varied from its minimum to maximum value. The hypothesis test compares the output curve for each simulation as the parameter is varied, against the original experimental data. The tipping point value is then studied further in the final detailed test. This has the effect of 'zooming' into the area between the dark and lightly shaded areas of the chart representing the h=0 (the null hypothesis, denoting that there is a match between the output curve for the simulation and the output curve of the experimental data) and h=1 (no match against the experimental data).

4.2.4. Results for Illness Duration Parameter with Population Size

52910

The results for illness duration variation for SD are shown in Figure 38 and for ABM, in Figure 39. This shows the variation of the range of output data achieved.

There is a significant peak at 5 weeks which also contains the largest IQR.
Figure 38: Output from the SD model for SIR with Population of 52910 and Illness Duration Variation

Figure 39: Output from the ABM model with Population of 52910 and Illness Duration Variation

The results of the hypothesis tests for SD are shown in Figure 40 which also indicates the matching region which is represented by a dark band followed by a tipping point followed by a non-matching region. For SD there is a 'tipping point' which occurs...
between values of 1.5 where the output matches the original data and 1.6 where it no longer matches.

The tipping point for ABM is 2.3 as shown in Figure 41 which represents the average tipping value over several simulations. As for SD, there is a distinct stepping point

![Figure 40: Hypothesis test for the SD model with Population 52910 and Illness Duration Variation](image)

**Figure 40: Hypothesis test for the SD model with Population 52910 and Illness Duration Variation**

The results of the detailed hypothesis tests for SD are shown in Figure 42. The ABM detailed hypothesis tests are also carried out and shown in Figure 43.

![Figure 41: Hypothesis test for ABM with Population 52910 and Illness duration Variation](image)

**Figure 41: Hypothesis test for ABM with Population 52910 and Illness duration Variation**
4. Variation of ABM and SD across Stepped Parameter Ranges

For SD, there is a matching region followed by a tipping point followed by a non-matching region. For ABM, there is no initial matching region. This is because at this detailed level, the tipping point is no longer confined within the range of values for the detailed simulation. Also, rather than a single tipping point there is a small band of matching and non-matching regions which is the result of different output curves.

In AnyLogic, for repeated simulations, initial seeding values are only set during the start of the simulation. This means that the actual seed value chosen for the random numbers for each simulation for the detailed analysis is different to that for the initial set of simulations.
4. Variation of ABM and SD across Stepped Parameter Ranges

4.2.5. Results for Illness Duration Parameter with Population Size 5291

The result of the experiments for illness duration variation for SD are shown in Figure 44. Each simulation took less than 0.1 seconds. The result for ABM is shown in Figure 45. Each simulation took about 0.45 seconds.

**Figure 44: Output from SD model with Population 5291 and Illness Duration Variation**

**Figure 45: Output from ABM with Population 5291 and Illness Duration Variation**
4. Variation of ABM and SD across Stepped Parameter Ranges

The results of the hypothesis test for SD with a population size of 5291 is shown in Figure 46. The results of the hypothesis test for ABM with a population size of 5291 is shown in Figure 47.

![Hypothesis test for SD model with Population 5291 and Illness Duration Variation](image1)

**Figure 46: Hypothesis test for SD model with Population 5291 and Illness Duration Variation**

![Hypothesis test for ABM with Population 5291 and Illness duration Variation](image2)

**Figure 47: Hypothesis test for ABM with Population 5291 and Illness duration Variation**

Both the SD and ABM outputs have distinct stepping points moving from matching to non-matching regions. The small additional band for SD may be the result of the application of the Wilcoxon rank sum test to slightly different shaped output curves.

The detailed hypothesis test results for the tipping point (the point where the test moves from non-matching to matching) for SD are shown in Figure 48.
Figure 48: Detailed hypothesis test for SD model with Population 5291 and Illness Duration Variation

The detailed hypothesis test results for the tipping point for ABM are shown in Figure 49.

Unlike the SD, the tipping point detailed values do not go from valid to non-valid. This is due to seed value chosen at detailed level which is also similar to Figure 43.

Figure 49: Detailed hypothesis test for ABM with Population 5291 and Illness Duration Variation

4.2.6. Results for Illness Duration Parameter with Population Size 529

The SD illness duration variation experiment for SD with a population of 529 is shown in Figure 50. The curve follows the same inter-quartile range profile as the full scale and one tenth scale experiment.
The illness duration variation experiment for ABM with a population of 529 is shown in Figure 51. Each single simulation in the experiment took 0.1 seconds.

![System Dynamics Model - Population 529 - Illness Duration Variation](image1)

**Figure 50: Output from SD model with Population 529 and Illness Duration Variation**

![Agent Based Model - Population 529 - Illness Duration Variation](image2)

**Figure 51: Output from ABM with Population 529 and Illness Duration Variation**

The SD hypothesis test is shown in Figure 52. The tipping point is consistent with the full scale population of 1.5. For this experiment, there is a short gap representing a non-matching region followed by a matching region. This is due to the random
interactions which generate different output curves. The hypothesis test for ABM with a population of 529 is shown in Figure 53 with a tipping point of 2.3.

Figure 52: Hypothesis test for SD model with Population 529 and Illness Duration Variation

Figure 53: Hypothesis test for ABM with Population 529 and Illness Duration Variation

The detailed hypothesis test results for the tipping point for SD are shown in Figure 54.

Figure 54: Hypothesis test results for the tipping point for SD
Figure 54: Detailed hypothesis test for SD model with Population 529 and Illness Duration Variation

The detailed hypothesis test results for the tipping point for ABM are shown in Figure 55. For the detailed hypothesis tests, there is a considerable set of matching and non-matching bands. This indicates that for detailed analysis, there is no uniquely identifiable tipping point at a detailed level.

This is due to a combination of local interactions generating different outputs combined with a smaller population set which results in a larger variation in the output results compared to the simulations at full scale.

Figure 55: Detailed hypothesis test for ABM with Population 529 and Illness Duration Variation

4.2.7. Total Variation for Illness Duration

The total variation (sum of differences across the inter-quartile range) is shown in Table 9:

<table>
<thead>
<tr>
<th>Population</th>
<th>Illness duration variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>52910</td>
<td>27263</td>
</tr>
<tr>
<td>5291</td>
<td>2257</td>
</tr>
<tr>
<td>529</td>
<td>273</td>
</tr>
</tbody>
</table>

Table 9: Total variation for Illness Duration for Agent Based Model of SIR
4.2.8. Results for Contact Rate Parameter with Population Size 52910

The result for contact rate variation for SD with population size 52910 is shown in Figure 56 and for ABM, in Figure 57.

**Figure 56: Output from SD model with Population 52910 and Contact Rate Variation**

**Figure 57: Output from ABM with Population 52910 and Contact Rate Variation**
The low level of variation observed in Figure 57 may be the result of the small world network combined with the contact rate variation. To understand the origins for this, further investigation would need to be taken but this is a useful indicator for any research interested in the effects of changes in the contact rate combined with the small world network. The SD hypothesis test is shown in Figure 58. The hypothesis test for ABM is shown in Figure 59.

**Figure 58: Hypothesis test for SD model with Population 52910 and Contact Rate Variation**

<table>
<thead>
<tr>
<th>Contact Rate Value</th>
<th>Hypothesis Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
</tr>
<tr>
<td>12</td>
<td>1.0</td>
</tr>
<tr>
<td>14</td>
<td>1.0</td>
</tr>
<tr>
<td>16</td>
<td>1.0</td>
</tr>
<tr>
<td>18</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Figure 59: Hypothesis test for ABM with Population 52910 and Contact Rate Variation**

<table>
<thead>
<tr>
<th>Contact Rate Value</th>
<th>Hypothesis Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
</tr>
<tr>
<td>12</td>
<td>1.0</td>
</tr>
<tr>
<td>14</td>
<td>1.0</td>
</tr>
<tr>
<td>16</td>
<td>1.0</td>
</tr>
<tr>
<td>18</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The detailed hypothesis test results for the tipping point for SD are shown in Figure 60. The detailed hypothesis test results for the tipping point for ABM are shown in Figure 61.
4. Variation of ABM and SD across Stepped Parameter Ranges

Figure 60: Detailed hypothesis test for SD model with Population 52910 and Contact Rate Variation

Figure 61: Detailed hypothesis test for ABM with Population 52910 and Contact Rate Variation

4.2.9. Results for Contact Rate Parameter with Population Size 5291

The result for contact rate variation for SD with population size 5291 is shown in Figure 62. The results show a peak level of infection between weeks 3 and 4 followed by a gradual decline in variation over the weeks.
The result for contact rate variation for ABM with population size 5291 is shown in Figure 63 showing a peak level of infection in weeks 4 and 5. Unlike the SD results, the IQR for the peak variation covers a larger range of values. For week 5, the minimum value for the IQR is 500 compared to 0 for ABM.

The SD hypothesis test is shown in Figure 64. The hypothesis test for ABM is shown in Figure 65.
Figure 65: Hypothesis test for ABM with Population 5291 and Contact Rate Variation

The detailed hypothesis test results for the tipping point for SD are shown in Figure 66. The detailed hypothesis test results for the tipping point for ABM are shown in Figure 67.

Figure 66: Detailed hypothesis test for SD model with Population 5291 and Contact Rate Variation

Figure 67: Detailed hypothesis test for ABM with Population 5291 and Contact Rate Variation
4.2.10. Results for Contact Rate Parameter with Population Size 529

The results for contact rate variation for SD with population size 529 is shown in Figure 68 and for ABM, in Figure 69.

**Figure 68: Output for SD model with Population 529 and Contact Rate Variation**

**Figure 69: Output for ABM with Population 529 and Contact Rate Variation**
The SD version shows peak infection during week 2 followed by a gradual decline in infection. For ABM, peak infection also occurs during week two but the IQR are much smaller compared to the SD version across the weeks. In addition, for ABM there is a much sharper rise in initial infections and a corresponding sharp decline in the infections compared to the SD version.

The SD hypothesis test is shown in Figure 70. The hypothesis test for ABM is shown in Figure 71.

Figure 70: Hypothesis test for SD model with Population 529 and Contact Rate Variation

Figure 71: Hypothesis test for ABM with Population 529 and Contact Rate Variation

The detailed hypothesis test results for the tipping point for SD are shown in Figure 72. The detailed hypothesis test results for the tipping point for ABM are shown in Figure 73.
4. Variation of ABM and SD across Stepped Parameter Ranges

4.2.11. Total Variation for Contact Rate

The total variation (sum of differences across the inter-quartile range) is shown in Table 10.

<table>
<thead>
<tr>
<th>Population</th>
<th>Contact rate variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>52910</td>
<td>54404</td>
</tr>
<tr>
<td>5291</td>
<td>5272</td>
</tr>
<tr>
<td>529</td>
<td>492</td>
</tr>
</tbody>
</table>

Table 10: Total variation for Contact Rate for Agent Based Model of SIR
4. Variation of ABM and SD across Stepped Parameter Ranges

4.2.12. Analysis of Results for Illness Duration and Contact Rate

It can be seen for the full scale experiments that the results show a non-matching region followed by a region with matching bands followed by an area which is non-matching. The reason for this is due to the random nature of ABM and the seeding value used. It is possible to set the seed value at the start of the simulation and predefined seeds are used to make the experiments reproducible for the initial match tests. However for the detailed tests, even though seed values are set, the random set of interaction is no longer matching up with the original main experiment. Therefore the interactions no longer occur in the same pattern as the original experiment and there isn't a region of matching data followed by a region of non-matching data.

For SD, the tipping point for the detailed test is consistent with that of the main SD tipping point experiment. There is no random number selection involved and therefore there is greater consistency in the results. This is always the case in simple ordinary differential equations as the equations are being solved and no random number generation is involved.

Compared to the tipping point in the full scale population of 52910 of 1.5, the tipping point lies in the same region when scaled down to population sizes of 5291 and 529.

SD scales easily from the full population to the scaled down population. The output curve at any scale for SD remains the same, however there is an initial time delay that needs to be taken into account at different scales. This has the effect of adjusting the position of the output curve so it can be matched up with the empirical data. Finding the warm-up time is more time consuming for ABM as each experiment takes longer.
and several simulations need to be run for each warm-up time considered so that the average can be taken.

For ABM at a population size of 529, the results show the values moving in sequence from matching against the empirical data to not-matching. The banding is seen more in smaller populations scales compared to larger population sizes. This is due to the larger amount of variation of the output curve seen in smaller population sizes.

Unlike the SD model, the profile of the output curve for ABM varies between population sizes both in terms of the inter-quartile range and the shape.

The experiments show how the variance compares between ABM and SD and some of the trade-offs between using SD and ABM in terms of performance and scaling.

SD is able to produce fast results and scales easily, producing consistent results between different population sizes in terms of the output curve of the simulations. In the SIR experiments, scaling for population sizes for SD was easily achieved using different warm-up times for the experiments. This can be useful where an aggregated overall simulation output is required to give a rapid indication of the average outcomes given specific input parameters.

Although SD is a simplification, most computer simulations are simplified representations of the real world and therefore SD may be suitable depending on the type of question involved in the study.

Basic SD alone is not able to produce variation in the output results but versions of it are able to do so [124].
In order to introduce variation for SD, parameters can be modified to determine the effect of different parameters on the output space of the simulation. However this does not realistically capture the variation that occurs at a person level, where the variation occurs due to the random nature of contacts between people in a given population.

To capture this fully, that interaction needs to be captured at a lower level, within the simulation itself, capturing the variation when it actually occurs. It is a form of interaction-based simulation which is also seen with simulations using bipartite graphs [125].

It is possible in ABM to generate different models with different types of network connectivity between people in a population to understand these scenarios. Having a model that is an accurate representation of the real-world is useful for the validation and verification of the model [112] and ABM helps to achieve this as it represents the real world more naturally.

One of the main issues with using ABM with currently available desktop computers is the time taken to run experiments. As the population size increases, the time taken to run the experiments increases rapidly. This is not the case with SD where the time taken to run the experiments does not depend on the population size. It is possible to scale the simulation down to different sizes for ABM. This could be done easily by modifying the warm-up time for the experiments. Scaling down the population size for ABM maintains the tipping point. However, detailed tipping point values are more difficult to obtain due to the way that random interaction is seeded for the simulation.
4. Variation of ABM and SD across Stepped Parameter Ranges

4.3. Conclusions

This chapter contained an implementation of SIR using SDM and ABM.

The experiments show that while SD can be a valuable form of obtaining average values for outcomes for SIR, ABM is able to produce the kind of outcomes associated with different types of interactions between infected individuals. It is able to do this from low levels of interaction due to low connectivity among individuals, to higher infectivity due to a larger set of interactions. Those range of values can help to determine the upper and lower boundaries of the spread of infection. Additionally it is possible to modify the connectivity among individuals in the population to play out what-if scenarios which can be used to help assess containment strategies for infectivity.

One of the main issues with ABM is the time taken to run experiments with large number of agents but it has been shown that ABM can be scaled down whilst maintaining the tipping point.

A number of studies have taken place to determine the best approach to the containment of influenza [126]. The aim of many of these studies is to try and understand the optimal intervention strategy to control the spread of infection and most cost-effective [127] and optimal approach to vaccination where required [128]. Computer simulation will continue to be a very useful tool in the decision making processes during influenza pandemics [129] and for producing guidelines for future intervention [130].

Although SD models and ABM models can produce similar results and SD continues to be a powerful and efficient tool for understand the spread of influenza, the local
level interactions that are part of ABM can be a useful feature. Firstly it can be used to understand what-if scenarios, for example to understand, in the event of an influenza outbreak, the effect of closing schools or major transport hubs. Secondly, it can give an insight into the low-level events, the interactions between individuals that cause the phenomena of emergence, in this case an outbreak of influenza. However ABM tends to work better with small to mid-range population sets.

For large populations (typically in the order of tens of thousands or millions of agents) performance becomes an issue making ABM an option that is no longer feasible. This research has shown that the scaling down of the population size of ABM can be used as one method to mitigate against some the issues related to ABM performance whilst preserving the natural variation generated by the random interactions between the individuals in the population and producing output that is equivalent to simulations based on SD.
5. Variation of ABM Output Compared to Real Variation

5.1. Introduction

The Monte Carlo experiment with SDM compared to ABM showed how ABM is able to capture variation in the output results. This leaves the question of the actual variation itself and how this relates to 'real' variation.

Real life variation is difficult to define. In the case of influenza for example, the variation can be different among different countries, cities, age groups and across different years and even within the same city, the dynamics of the variation can differ depending on the type of infection [131]. This is because the dynamics of interactions within any population is affected by many unknown internal factors (such as the contact links) as well as many external factors such as migration and transport links into and out of the population. Therefore if we are to compare the variation of a model against the variation of a real world system, it needs to be within a specific and well defined context.

In this experiment, ABM Model is developed using the influenza profile of a selected region within the United States based on publicly available influenza data. The parameters for the model are estimated and fine-tuned and variation from the output from the model is compared to the variation found in the empirical data. This gives information is used to evaluate the level and the profile of the variation from the ABM Model.
Influenza data has been collated for many years but following the 2009 Influenza A H1N1 pandemic there has been a gradual increase in the collection of influenza data. This has been in the form of official records from clinics and hospital admissions [132] but also a number of novel data collection approaches including the use of Twitter chats [133] and influenza search activity [134]. The results of these novel approaches appear to tie fairly closely with real influenza data although a little delay is involved.

The sections are laid out as follows:

- The influenza data and its source is described
- An ABM Model is built based for a single region
- Charts are presented for the variation of the influenza data compared to the variation of the ABM Model
- Charts are presented showing the effect of changes in the environment of the ABM Model on the variation
- An analysis and discussion summarises the main findings of the experiment

5.1.1. The Source of the Influenza Data

In the United States, The Centres for Disease Control and Prevention (CDC) make influenza data available to the public [135]. The data is presented in the form of a weekly influenza surveillance report known as FluView which contains key indicators for the follow.

This is made available in an online dashboard which allows the data to be viewed in the form of a map where the United States is split into different areas based on Census Divisions, HHS (Health and Human Services), and National levels. The data is
collected from the World Health Organisation (WHO) and the National Respiratory and Enteric Virus Surveillance System (NREVSS). The data is collected from health departments and healthcare providers. In the initial collection in 1997 there were approximately 250 providers of influenza data which has now increased to over 3000 providers.

5.1.2. Influenza Area

For the purpose of generating the ABM Model, data from a census division was chosen.

The diagram in **Figure 74** shows the different census divisions for the United States taken from U.S. Energy Information Administration, Office of Energy Analysis.

![United States Census Divisions](image)

**Figure 74: United States Census Divisions from U.S. Energy Information Administration, Office of Energy Analysis**

The New England Census division comprises:
5. Variation of ABM Output Compared to Real Variation

- Connecticut
- Maine
- Massachusetts
- New Hampshire
- Rhode Island
- Vermont

New England was chosen as it has a relatively lower population of approximately 14 million (based on the 2010 census) compared to the other census divisions.

5.1.3. Influenza Data

Influenza data is available from as early as 1997 but data for the H1N1 variation is only available since 2008. To avoid a bias in the results, only the influenza data available from this year onwards is used so that every year has a consistent set outputs based on a defined set of influenza variants.

Influenza variants from 2008 to 2014 are shown in Table 11:

<table>
<thead>
<tr>
<th>Influenza Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>A(H1)</td>
</tr>
<tr>
<td>A(H3)</td>
</tr>
<tr>
<td>2009 H1N1</td>
</tr>
<tr>
<td>B</td>
</tr>
</tbody>
</table>

Table 11: Influenza variants from CDC data for 2008 to 2014
5. Variation of ABM Output Compared to Real Variation

5.1.4. Real Influenza Data for 2008-2009

The data and chart images for the influenza data shown here is collected from the FluView online portal which allows users to retrieve flu data and associated charts by season and geographic area [136].

![Influenza data for 2008-2009](image)

**Figure 75: Influenza data for 2008-2009**

The data for 2008-2009 contains two waves of peak infection in winter 2009 with a value of 345 persons and in summer of 2009 with a value of 746 persons as shown in Figure 74.

This is characteristic of many types of influenza and there have been studies to predict these kinds of patterns [137] and is of significance in pandemics such as the 1918 influenza with mortality rates of approximately 50 million worldwide [138]. The 1918 influenza came in 3 waves across two years, spring 1918, autumn 1918 and spring 1919.
5. Variation of ABM Output Compared to Real Variation

Rather than treating this data as two independent simulations, the data for both of these curves has been taken as one set. This is so that a comparable year to year analysis of the variation can be made without introducing any bias in terms of selection of data.

5.1.5. Real Influenza Data for 2009-2010

Figure 76: Influenza data for 2009-2010

The influenza data for 2009-2010 contains a single peak of 794 persons in late 2009 as shown in Figure 76.

The characteristic is similar to a standard SIR profile except that there isn't a sudden surge of infections followed by a gradual decline. The profile is closer to a normal distribution. This may be due to the level of infectivity occurring within the population earlier in the year. The influenza curve follows closely after the previous two waves of influenza which occurred in 2008-2009. This represents three waves of influenza activity followed by a gradual decline in infections in 2010.
5. Variation of ABM Output Compared to Real Variation

5.1.6. Real Influenza Data for 2010-2011

The 2010-2011 influenza data also has the bell curve profile which peaks at 356 persons infected instead of the characteristic SIR curve of high initial infectivity followed by a gradual decline in infectivity due to immunity from the disease. This is shown in Figure 77.

The spread of data is significant in this curve. In the 2009-2010 data the spread of peak infectivity is 8 weeks compared to 13 weeks for the 2010-2011 influenza data, taking into consideration all weeks where infectivity was more than 100 persons. One explanation for this is that the average illness duration of the infection is larger in the 2010-2011 data compared to the previous year. Another explanation is that there were a number of different variants of influenza in 2010-2011 compared to the previous year where the primary variant was H1N1.

Figure 77: Influenza data for 2010-2011

The 2010-2011 influenza data also has the bell curve profile which peaks at 356 persons infected instead of the characteristic SIR curve of high initial infectivity followed by a gradual decline in infectivity due to immunity from the disease. This is shown in Figure 77.
5.1.7. Real Influenza Data for 2011-2012

The data for 2011-2012 is shown in Figure 78 and has the profile of a single curve with a gradual increase in infectivity caused by influenza A(H3) followed by a rapid drop in the infected population. This profile is the opposite of 'typical' infectivity where the initial infection is rapid as it gets passed from one person to another within the population and then there is a slow decline as people recover from the infection and no longer pass it to others. In this chart, the decline is more rapid and not gradual.

Although the major contributor of infection is influenza A(H3), there is also infection caused by influenza A, B and 2009 H1N1 but this is at comparatively lower levels compared to A(H3). Compared to the influenza data for 2010-2011, influenza 2009 H1N1 did not play a significant role in the total number of infections and the contribution of the total number of infections is shared between influenza variants A, B and 2009 H1N1.
5.1.8. Real Influenza Data for 2012-2013

The data for 2012-2013 is shown in Figure 79 and follows a more 'classic' SIR curve [139] as shown in Figure 80 with a relatively rapid increase in infectivity during winter 2012, peaking at 462 persons followed by a gradual decrease in the total number infected throughout the year of 2013.

Figure 79: Influenza data for 2012-2013

Figure 80: Aggregate Death Rates for US Regions for 918-1919 Influenza Pandemic from [139]
5. Variation of ABM Output Compared to Real Variation

5.1.9. Real Influenza Data for 2013-2014

The 2013-2014 influenza occurred in two waves during winter and spring of 2014 peaking at 296 persons during winter and 185 persons during spring. However unlike the 2008-2009 profile where one peak complete and the other started, in the 2012-2014 profile the two waves of influenza infectivity overlap. This is shown in Figure 81.

The reason for this is that in 2008 there was infectivity due to influenza A(H1), A(without subtype) and B between the weeks of 1 and 16 which largely recovered by week 16. This was then followed by influenza type 2009 H1N1 from weeks 17 onwards.

In contrast, in the 2013-2014 influenza profile, there is an initial wave of infectivity of 2009 H1N1 but mid-way during this way there is an outbreak of influenza A(H3). So the initial wave has not reached recovery stage and therefore we see the overlap.
5. Variation of ABM Output Compared to Real Variation

5.1.10. Parameters used for the Experiments

The ABM Model built for the study uses the three states of Susceptible, Infectious and Recovered as shown in Figure 82.

![Agent Based Model for analysing influenza variation](image)

For this model, the following parameters are required:

- Contact rate
- Probability of Infection
- Illness Duration (days)
- Total population

One additional element of the ABM Model is the environment. For this, a small world network topology is used as it represents local clusters with connections to other clusters. The space is set to be continuous 2D. The environment parameters are obtained using repeated simulations to select the parameters which fit the data as this information is not available from clinical records. The small world parameters set to the values shown in Table 12.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connections per agent</td>
<td>20</td>
</tr>
<tr>
<td>Neighbour link fraction</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 12: Network parameters for the Agent Based Model
From clinic, records the population size used is 17071 persons based on total admissions for the 2008-2014.

Having obtained the total population size, the estimation for the remaining parameters is required.

A common challenge when dealing with influenza modelling is the need to estimate the potential for transmission of the infection among individuals in the population with respect to real world statistics. One method for producing the estimation is the use of Monte Carlo Methods to obtain the parameters which can then be fed into a computer simulation [140].

This requires multiple cycles whereby a set of parameters are randomly selected from a normal distribution and the simulation is run. Repeated runs lead to different output curves. An average output needs to be taken for each time unit for the curve to compare against the raw data which is the influenza data from the CDC in this instance. The number of simulations required varies depending on the type of model involved and may in itself require an initial experiment to determine the number of simulations required to obtain a smooth output.

The different methods of stochastic analysis in Anylogic include:

- Optimisation using the OptQuest optimiser
- Sensitivity Analysis using overlaid charts
- Monte Carlo simulation experiments
- Calibration experiments

A summary of the process involved in fitting the data is shown in Figure 83.
Figure 83: Strategy for determining model parameters

The experiment using the OptQuest optimiser [141] is able to run replications which are repeated experiments using the same parameter values which are then aggregated in order to decide the next parameters to use. The user interface automatically shows the results of the experiment and the best solutions found. The experiment can be very
time consuming, depending on the number of parameters involved in the model but the automated aspect of the experiment means that it can be left unattended until the best solution is achieved.

Once the estimates are obtain, it is possible to further fine-tune the input parameters based on knowledge of how each input parameter affects the output. For example, if the illness duration is increased then the width of the output curve tends to be larger and vice versa. If the contact rate is increased, this leads to a more started gradient towards peak infection.

Although the total population is identified as 17071 persons, for the purpose of these experiments, this figure is reduced by a factor of 10 to 1707 persons. This is due to the large number of runs required for the initial Monte Carlo simulation to find the approximate parameters, further fine-tuning of parameters and the remaining experiments which also include 1000 simulation runs. The results that are shown are then scaled up by a factor of 10 so that comparisons can be made between the simulation data and the influenza data. In a scenario where model parameters are available and don't require estimation via simulation, it would be more suitable to use the full population set where feasible, depending on the simulation time.

Using this method, the values shown in Table 13 are obtained for the simulation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact rate</td>
<td>3.9</td>
</tr>
<tr>
<td>Probability of infection</td>
<td>40</td>
</tr>
<tr>
<td>Illness duration (days)</td>
<td>1.6</td>
</tr>
<tr>
<td>Total population</td>
<td>1707</td>
</tr>
</tbody>
</table>
Table 13: Input parameters for the Agent Based Model

Using these parameters, a total of 6 simulation runs are performed based on the model. This is so that the 6 years of real data can be compared against the relative 6 runs of ABM simulation output. An alternative is to run the ABM simulation over 100 or 1000 successions and use the averages of those simulations for comparison. This has also been done in the second stage of the experiment, however the purpose of the first stage of the study is to compare a 'typical' output from the simulation to a 'typical' years worth of influenza data. Influenza output curves can vary considerably from year to year as shown in the discussion of influenza data. The output of ABM simulation can also vary considerably so this makes the comparison like-for-like.

In the second stage, 1000 simulation runs are performed. These are performed to give an account of the projected range of values covered by the ABM Model. Although the equivalent information is not available from historical influenza data, it does provide an estimation of the boundary of space for which a future influenza outbreak may cover and also helps to understand the behaviour of the ABM Model for large simulation runs.

5.2. Results for Variation of ABM Compared to Real Data

The results for both the influenza data from the Centers of Disease Control and the ABM simulation is presented using box plots for each of the weeks. MATLAB is used to generate the box plots which show for each week:

- The minimum value
- The maximum value
- The median
5. Variation of ABM Output Compared to Real Variation

- The inter-quartile range
- Any outliers identified

Although the charts show the outliers, these are still useful for analysis because they represent real values for the influenza data over a period of 6 years worth of data or 6 simulation runs. As the number of years or runs is low, for these cannot be considered as real outliers, this would require a much larger number of runs.

5.2.1. Statistical aspects of real influenza data

The forecasting of influenza is a difficult task due to the number of assumptions that are made due to the difficulty with which to obtain specific formal knowledge surrounding the infectivity.

Statistical analysis of influenza is also difficult due to the cyclical and irregular nature of the infectivity [142]. Other influences include the weather, with the infection being more prevalent in winter months as well as viral mutations and outbreaks outside the area of surveillance [143].

The ABM Model of influenza used in the experiments in the chapter always produce a single output curve whereas real data can produce multiple output curves of irregular patterns. Two approaches can be considering for overcoming this limitation:

- Generation of an SD model of influenza containing irregular waves of infectivity together with an equivalent version using ABM.
- Breakdown of the waves of influenza to single output curves so a direct comparison of a single output curve from the ABM Simulation can be compared with a single curve from real data.
The difficulty with the first approach, the modelling of multiple waves, is that the role of re-infection in the generation of multiple waves is poorly understood and the resulting model becomes difficult to validate [144]. As a consequence, the modelling of multiple waves using an ABM Model becomes equally difficult as knowledge of the causes of re-infection need to be identified and implemented in the model. The resulting System Dynamic and ABM Models in these circumstances, whilst generating output profiles that may match some instances of real data would be non-validated and consequently limited for further research (to determine cause and effect of parameter changes for example).

The second approach requires breaking multiple waves of infectivity down to separate waves so each wave can be compared with a single curve from an ABM Model. The difficulty with this approach is that the waves of infectivity have no clear starting and finishing point as the curves are irregular. For example, in Figure 81: Influenza data for 2013-2014, the curve for the total number of infections for influenza virus type A(H3) has no clearly defined range of wave start and end points. Visually, it is possible to identify three distinct waves from week 44 from 2013 to week 5 in 2014, week 5 to week 7 in 2014 followed by a final wave from week 7 to week 21 in 2014. Additionally, this could also be defined as a single wave for the entire period of data. Isolating such waves is therefore subjective and the identified waves are not validated making any comparisons between an ABM Model and SDM model subjective.
5.2.2. Infected Population Based on Real Data

The infected population values based on real data from the Centers of Disease Control is shown in Figure 84 in the form of box plots covering each of the 6 years across the different weeks.

![Figure 84: Infected population for New England 2008-2014](image)

The total number of infections across the inter-quartile range is 3965 persons. Although there are waves of infections over the time period over the 6 years, the average peak infections for the real data occur between weeks 10 and 20.

5.2.3. Infected Population using an Agent Based Model with 6 Runs

The infected population values based on the ABM Model is shown in Figure 85.
Variation of ABM Output Compared to Real Variation

Figure 85: Infected population using an Agent Based Model with 6 runs

The total number of infections across the inter-quartile range is 2960 persons. Peak infections for the simulation output occurs between weeks 9 and 20 which is very similar to the results for the real data where the peak numbers of infections for the population occur between weeks 10 and 20. Like the real data, the numbers of infections reduce slowly over time but in the simulation output the immunity takes longer over the simulations.

Although the results from the ABM Model are similar to the real data, the real data can contain multiple waves of infectivity compared to the result from the ABM Model which will always have a single wave of output. Nevertheless, the results are a useful indication of the spread of variation over time and shows that, in terms of variation, there can exist more than one region where peak levels of variation occur. As variation is the main theme, this can be useful for the purpose of research.
5.2.4. Infected Population using an Agent Based Model with 1000 Runs

Although the variation experiments in this chapter do a like-for-like comparison using 6 years worth of real data against 6 simulation outputs, the experiment is repeated using 1000 simulation runs to show the effect of increasing the number of simulations.

![Infected Population Using Agent Based Model with 1000 runs](image)

**Figure 86: Infected population using an Agent Based Model with 1000 runs**

The total number of infections across the inter-quartile range is 2960 persons.

5.2.5. Comparison of Averages for Influenza Data from CDC and ABM

The median values for the influenza data from the Centers of Disease Control and for the output from the simulations from the ABM Model is shown in Figure 87.
5. Variation of ABM Output Compared to Real Variation

Figure 87: Median values from influenza data from CDC and Agent Based Model

The medians correspond to the medians shown in the box plots for the two independent experiments but here, they are overlaid so they can be compared more easily. The statistics for the output curves are shown in Table 14.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>CDC</th>
<th>ABM (6 runs)</th>
<th>ABM (1000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>157</td>
<td>205</td>
<td>140</td>
</tr>
<tr>
<td>Mean</td>
<td>39</td>
<td>62</td>
<td>72</td>
</tr>
<tr>
<td>Median</td>
<td>11.25</td>
<td>25</td>
<td>80</td>
</tr>
<tr>
<td>Range</td>
<td>157</td>
<td>205</td>
<td>140</td>
</tr>
</tbody>
</table>

Table 14: Median Statistics for CDC and ABM curves
5.2.6. Impact of the Number of Connections per Agent Overview

These charts show how the variation is affected by changes in the connections per agent. These are grouped as

- Plots for the infected population for each week as connection is varied
- The IQRs for each week as connection is varied
- The sum of interquartile variation for each parameter as connection is varied

5.2.7. Infected Population as Number of Connections are Varied

The box plots of the infected population are shown in Figure 89 showing the ranges including any outliers identified by MATLAB, the IQRs and medians.
5. Variation of ABM Output Compared to Real Variation

- Agent Based Model with 10 connections
- Agent Based Model with 20 connections
- Agent Based Model with 30 connections
- Agent Based Model with 40 connections
- Agent Based Model with 50 connections
- Agent Based Model with 60 connections
5. Variation of ABM Output Compared to Real Variation

5.2.8. Sum of Interquartile Range as Number of Connections are Varied

The sum of the IQR for each of the connections is shown in Table 15.

<table>
<thead>
<tr>
<th>Connections</th>
<th>Sum of IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3670</td>
</tr>
<tr>
<td>20</td>
<td>5220</td>
</tr>
<tr>
<td>30</td>
<td>3640</td>
</tr>
<tr>
<td>40</td>
<td>4500</td>
</tr>
<tr>
<td>50</td>
<td>6830</td>
</tr>
</tbody>
</table>

Figure 89: Infected Population as connections per agent are varied
5. Variation of ABM Output Compared to Real Variation

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>9610</td>
</tr>
<tr>
<td>70</td>
<td>8850</td>
</tr>
<tr>
<td>80</td>
<td>6550</td>
</tr>
<tr>
<td>90</td>
<td>5900</td>
</tr>
<tr>
<td>100</td>
<td>9150</td>
</tr>
</tbody>
</table>

Table 15: Sum of interquartile range as connections are varied

5.2.9. Interquartile Range as Number of Connections are Varied

The value of the IQR for each of the weeks as the connections are varied from 10 to 100 in steps of 10 is shown in Figure 90.
Figure 90: Interquartile range as connections per agent are varied
5. Variation of ABM Output Compared to Real Variation

5.2.10. Impact of the Neighbour Link Fraction

The effect of neighbourhood link fraction changes are shown using in Figure 91 using box plots for each week, and the IQRs for each week.
5. Variation of ABM Output Compared to Real Variation

![Figure 91: Infected Population as neighbour link fraction is varied](image)

These charts show how the variation is affected by changes in the neighbour link fraction which represents the percentage of links which are rewired to other random agents.

### 5.2.11. Interquartile Range as the Neighbourhood Link Fraction is Varied

The values for the IQR for each of the values for the neighbourhood link fraction from 0.1 to 1.0 in increments of 0.1 is shown in Figure 92.
The data from these charts is also shown as part of the box plots in Figure 91 but in Figure 92, only the medians for the IQR is shown so that the median values can be easily compared for the different values for the link fraction.
Figure 92: Interquartile range as neighbourhood link fraction is varied

5.2.12. Sum of Interquartile Variation

The sum of the IQR for each of the values for the IQR is shown in Table 16.

<table>
<thead>
<tr>
<th>Neighbour link fraction</th>
<th>Sum of IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>2730</td>
</tr>
<tr>
<td>0.2</td>
<td>5220</td>
</tr>
<tr>
<td>0.3</td>
<td>5840</td>
</tr>
<tr>
<td>0.4</td>
<td>3690</td>
</tr>
<tr>
<td>0.5</td>
<td>4630</td>
</tr>
<tr>
<td>0.6</td>
<td>4330</td>
</tr>
</tbody>
</table>
5. Variation of ABM Output Compared to Real Variation

<table>
<thead>
<tr>
<th>Neighbour Link Fraction</th>
<th>Sum of Interquartile Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>2740</td>
</tr>
<tr>
<td>0.8</td>
<td>4340</td>
</tr>
<tr>
<td>0.9</td>
<td>1730</td>
</tr>
<tr>
<td>1.0</td>
<td>480</td>
</tr>
</tbody>
</table>

Table 16: Sum of interquartile range as neighbour link fraction is varied

5.3. Analysis of Results for Variation of ABM Compared to Real Data

In this section, the results are analysed and conclusions follow in the subsequent section of this document.

5.3.1. Box Plots for Infectivity

The influenza data from the Centers of Disease Control contain two types of output curves. The first type has a single peak and the second type has two peaks which represent the two waves of influenza infectivity. As a result of this, in the box plots for the influenza data, two sets of output curves are visible in the chart. For the IQR, there is initially a wider range of results for each of the initial weeks of infectivity, from week 10 to the week 23. This is then followed by a gradual decline in which the range of values for the inter quartile range reduces from week 23 to week 35. The IQR is useful for the analysis as it represents the spread of the central half the data and insensitive to outliers.

One of the reasons why there is a larger band of IQRs shown in the box plots earlier in the infection is due to the shape of the infection curve in the CDC influenza data. In the classic SIR representation, there is a surge of infection in typically all cases of infection. This leads to a sharp gradient early on in the infection. As this applies to
most infections, the IQRs early on are relatively low compared to the rest of the charts where the range of values depends more on factors such as the length of illness. The factors which affect this are depend on:

- The infections start at different weeks through the year
- The infection can occur in waves with overlapping and non overlapping waves
- The shape of the infection curve varies from rapid infectivity to gradual infectivity
- The influenza data represents the results of several different types on influenza
- There is migration into and out of communities
- There are varying levels of initial infection (this can occur due to simultaneous infections across population or a single infection brought about by an existing member of the population or a new infected person entering the population)
- There is an assumption in the ABM Model that those infected recover but it is likely that several people do not fully recover and it is also likely that several deaths occur over the population
- Births and deaths (other than influenza) alter the population counts

The experiment based on the ABM Model of 1000 simulation runs shows a smooth distribution of IQR. The IQR is largest between weeks 15 and 35 or the central portion of the simulation time. This coincides with the common expected profile of the basic version of the SIR model where there is a small number of infections at the beginning and recovery at the end.
5.3.2. Medians for Influenza Data from CDC and ABM Data

The influenza data from the simulation built on the ABM Model over 6 runs has two main peaks at week 13 and week 28. As each simulation generates a different output curve, this represents the two peaks from those simulation outputs. Unlike the IQR from the CDC, the IQR from the simulation is more uniform whereas the CDC IQR has sharp peaks and troughs. This implies that even for small sample sizes, the amount of variation exhibited by real influenza data comprising several variants of influenza is slightly more unpredictable than a like-for-like sampling from an equivalent ABM Model. It is likely that any small sample size of about 6 years will produce different results.

Over time this is likely to be averaged as the sample size increases as shown in the simulation with 1000 runs. This is based on the law of large numbers, the branch of probability theory that states that over a large number of repeated trials, results will become closer to the expected results [145]. In this instance we have no way of knowing the actual expected result as there are too much random aspects to influenza (such as migration and mutations that change the dynamics of the infection) to make an accurate picture of what the results would be. However, with the ABM Model of over 1000 runs, what we do obtain is the most likely range of that expected value based on the environment and the parameters used in the simulation.

The data from the statistics table suggests that the peak level of infection drops as the number of simulations is increased. Initially, the peak value for the infections is 205 for the 6 simulation runs which is slightly more than the 157 infections from the influenza data from CDC. This drops to 140 for the average over 1000 simulations.
This is due to the larger sample set showing again the significance of the number of simulations.

5.3.3. Interquartile Ranges for CDC and ABM Data

The charts for the IQRs show the spread of the ranges for each of the 3 types of simulation. The influenza data from the Centers of Disease Control show sharper spikes compared to the simulation from the ABM Model broadly in line with the medians. The IQR for the ABM Model has a good overall fit although there are no large spikes considering the small sample size of 6 years worth of influenza data and the output from 6 simulations.

The IQR for the simulation with 1000 runs is much smoother and regular in shape due to the averaging of a large number of simulations. If the number of simulations was extended to a million runs, this curve would have the same profile but would be much smoother due to the averages obtained over a large simulation set.

The IQR for the 1000 simulations shows that as the number of simulations is increased, the level of randomness or sporadic values in the output for the simulation is reduced in line with the law of large numbers in probability theory. Unlike the 6 years worth of data and the 6 simulations, there is no full recovery of infection for the 1000 simulations. If the simulation time was extended then this would even off as full recovery in the population set takes place. This represents the scenario where the length of recovery may take considerably longer than a typical cycle of recovery in an influenza infection where full recovery takes place before week 50.
5.3.4. Connections Stepping Experiment

The connections experiment showed how the change in the variation of the output across the different weeks as the number of connections is stepped from 10 to 100 connections per agent. The results are presented as box plots a table with a summary of the total IQR for the different values of the number of connections.

As the number of connections is increased the following is observed:

- There is no significant increase in the peak infection value when the connection is between 10 and 40 and no significant change in the variation of the output.
- Connection values between 50 and 100 have a greater peak infectivity value but there is no linear correlation between these values. There is also no significant correlation between the total sum of variation represented by the IQR and the number of connections within this range.
- The profile of the variation is irregular in line with the variation observed in the real data. For example with 60 connections we see three distinct waves of infectivity. There is an initial wave up to the first 10 weeks followed by another wave of infectivity between weeks 10 and 25 followed by a third wave for the remainder of the time.
- The largest amount of variation is observed when the number of connections is 60. The smallest amount of variation is observed when the number of connections is 30.
5. Variation of ABM Output Compared to Real Variation

5.3.5. Neighbourhood Stepping Experiment

The neighbourhood experiments showed the effect of changes in the neighbourhood on the peak infectivity and the variation which is represented by the IQRs.

The following is observed as the neighbour link fraction is increased:

- Peak infectivity occurs between the values of 0.2 and 0.6
- Infectivity is highest at the value of 0.2
- There is a sharp dip in the level of infectivity after 0.7
- Peak infectivity and total variation is least for the values of 0.9

The results of the experiment suggest that peak infectivity and variation occurs for lower levels of the neighbour link fraction. This is due to the strong local clusters of connections for the smaller values as shown in Figure 93.

![Figure 93: Representation of local clustering in an Agent Based Model](image-url)
As the neighbour link fraction is increased, local clustering is reduced. This means that it is no longer possible to have a surge in infectivity as there are fewer clusters of people that are likely to spread infection quicker.

For the neighbour link fraction of 1.0, this represents a fully random network topology with no predefined local clusters so in this instance, the network topology has changed significantly compared to that for when the neighbour link fraction is 0.1 to 0.2.

Although there are a number of influenza models, the nature and complexity of influenza and the problems of ascertaining real data mean that sometimes it is necessary to use hypothetical communities in simulation [146]. However, information can be very useful even if aggregated at higher levels in abstract models as the model is then based on more realistic assumptions.

### 5.4. Conclusions

This experiment examined the variance generated by an ABM Model of SIR compared to the variance of real influenza data from the CDC. The specific research is based around the ability of a simulation from an ABM Model to generate a level of variation which can be fitted to the level of variation exhibited in real world data.

Based on the 6 box plots and the charts for the median and IQRs for the 6 sets of influenza data and simulation output, we can conclude that the results from the ABM Mode show that the simulations from the model provide a comparable level of variation for the 6 sample sets of influenza data from the Centers of Disease Control. This variation is generated internally, caused by the random contacts within the
population causing transmission of infection and not by the repeated modification of the input parameters for the model.

A number of assumptions have been made due to limited data available. These include the following:

- The infection is imitated from a single initial infected person
- The population size is static
- The contact structure is based on a small world network

With simulations based on real world data, there is often some kind of abstraction made either due to the unavailability of the necessary data or due to other constraints such as computation time and complexity which can occur as a result of introducing finer levels of detail. Much of this is based on the specific question that needs to be answered using the model. The model in this instance is developed specifically to answer the question of variability.

In the experiments carried out in this chapter, the real data comprises an amalgamated value for the infected population for five different types of influenza over time. It is also possible to select a single type of influenza with which to compare results. However, different types of influenza have different output profiles over time. For example in some years, there may be very minimal levels of infections compare to other years which may have considerable levels for the infected population. Comparing against all five types of influenza therefore minimises these effects to give a more generalised view of infectivity and reduces the side effects of issues involved using specific influenza types over time.
5. Variation of ABM Output Compared to Real Variation

The model parameters for the infection have been estimated using Monte-Carlo simulation to determine the parameters. In real world systems, parameters can be difficult to estimate. As model abstraction increases, additional problems are introduced as these parameters are aggregated.

Estimation of parameters for ABM Models has significance performance impacts when used in a desktop environment with standard hardware. This is a result of the need to perform large numbers of simulations in order to extract the averages which can then be used to fit against the real life data. If variation becomes another level of data to fit then this can add an additional performance impact. Therefore, a reduction in the population size is sometimes necessary in order to reduce the overall time constraints for the project.

The results for the 1000 simulation runs shows that the level of variation for larger quantities tends to become more uniform, following the shape similar to the normal distribution as expected with the larger sample set.

A single output from a simulation based on an ABM Model generates a 'typical' output for influenza over any year. In much the same way that the real world data varies significantly from year to year, a repeat of the simulation using random seeding generates a different output. On its own, a single simulation can only give a typical output curve but several small numbers of output curves can be very useful if fitted against empirical data as they are then able to offer a useful spread of variation.

If only a single set of empirical influenza data is available then determining the output variation from the actual data can be difficult to determine from the data alone but it may be possible to estimate the variation based on existing mathematical knowledge.
For example if we know the mix of contact rate within a population then this can be used to draw conclusions about the spread of variation by stepping the contact rate from the lower value to the upper value based on a normal distribution.

The generation of an ABM Model is a complex process where parameters are not fully understood due to the length of time taken to estimate the parameters. For the purpose of this experiment, this has been mitigated by constraining the model to a smaller dataset (covering the region of New England) and limiting the number of datasets used for analysis (to 6 years of influenza data) and further simplifying the actual ABM Model itself by reducing the population size.

The level of simplification of the model depends on the nature of the question so if the question was based around a study for the entire population of the United States over a defined period of years then this may not have been a suitable approach. Therefore for ABM, there isn't a simple set of rules for increasing the efficiency of the simulation but a useful set of guidelines can be used to improve efficiency based on the specific scenario and question.

The changes in the number of connections suggest that the total level of variation is not significantly affected by connections of up to 40 agents. Beyond this, there is a rise in the number of infections and also the level of variation. More importantly, based on the small world network topology with the fixed neighbourhood link fraction this experiment shows that beyond a certain number of connections (in this case 40), there is no significant change in the infectivity or range as the number of connections is increased further.
5. Variation of ABM Output Compared to Real Variation

This is of significance where a fit of the variation is required with real data as this implies that a range of values for the number of connections can all generate a similar quantity of variation and peak infectivity.

The results from the stepping experiments for the neighbour link fraction suggest that there is an optimal range of values for the neighbour link fraction for peak infectivity. In the experiments, this is found to be in the range of 0.2 to 0.6. This range sees the largest values for infection as well as the largest variation in the results. Infectivity and variation becomes very low when the network topology becomes random.

This suggests that a small world network topology is effective at transmission of infection when there are local clusters of people and when between 20 to 60% of the connections are rewired to other people. It is found that random networks are not very effective at infection transmission compared to small world networks.

One conclusion that can be drawn from this is that the range of variation depends on the network topology of the connections. Standard SDM models assume an even distribution of connectivity similar to a homogenous view of the random network topology. This does not contain locally connected clusters of people within the population. Therefore the dynamics of the variation is different. As seen from the first experiments using SDM with Monte Carlo simulation, the range of variation in the output space is smoothly distributed and has direct correlation with the peak infectivity. The range of variation in the output space observed in the ABM Model is a result of the transmission of infection within and across clusters of people in the population.
The Standard SDM model has no internal concept of network topology. Therefore variation can only be introduced through parameter modification. The ABM Model is able to introduce variation in the output space by manipulation of the network structure whilst keeping the input parameters intact. This can be very useful because it allows a research to be able to understand the impact of changes in the connections projected ranges of values in the output space. For example, if the small world network topology is used to represent local communities with the 'rewiring' of local links representing public transport then one of the things that can be estimated is the effect of reducing the levels of public transport by reducing the neighbour link fraction.

The experiments show that the total level of variation is closely related to the peak infection values and these tend to have a stepped correlation rather than a linear correlation. The results of the experiments suggest that only significant reductions in the neighbour link fraction have any noticeable impact on peak infectivity and that there is greater variation in the output space for lower values of the neighbour link fraction. Based on the experiment parameters, this would therefore tend to imply that a reduction in public transport may have little effect unless it is reduced by a very significant level and that there is a greater spread of infectivity for when public transport is reduced.

The level of peak infectivity and variation in the output space is different for each simulation based on an ABM Model. A question arises about the number of simulations which are necessary for an experiment. There is limited research which aims to study aspects of variation and number of simulations. We know from the probability of large numbers that if a simulation is repeated a large number of times of
e.g. 1,000 simulation as per the experiments in this section in the case of SIR, that this produces a smooth output curve when the averages are used to plot the output space. This also has an impact on the range which generates more uniform bands.

Building computer models can be difficult. It can be perceived that building ABM Models can be more difficult because, as well as generating multiple outputs to take the average, those number of simulations must be determined and due to the extra parameters related to the environment, those must also be determined. Along with obtaining local level knowledge required for the interactions, creating credible ABM Simulations can be challenging [147].

The number of simulations therefore depends on the nature of the research or study. In initial experiments, 6 years worth of data is compared to 6 runs of simulation. This generates a typical output and range across 6 years and therefore can be considered as a representative projection based on 6 years worth of data. Had further data been available for e.g. 50 years then it may be more feasible to generate 50 simulations for comparison and fitting.
6. Conclusions

6.1. Introduction

This chapter concludes with a summary of the research findings, the contribution to existing knowledge, limitations of the research presented in this thesis and a discussion of possible further work.

Early opinions for mathematical modelling of epidemic models were that deterministic models gave an average outcome of a corresponding stochastic model and that for large populations, it was the average that mattered. A more recent understanding is that both deterministic and stochastic models have their strengths and weaknesses.

Traditional deterministic models of epidemiology assume heterogeneity of mixing. It is assumed that individuals have the same rate of contact with others and recovery from infection takes the same time. In reality, contact rate is affected by transport networks and individual lifestyle and recovery from infection can depend on age and other factors so these are not take into account. Sometimes this data is difficult to ascertain but in smaller population sizes it may be possible to obtain this information and build a model that is a closer representation to reality.

One of the underlying reasons why epidemiological systems exhibit variation is due to the complex way that the individuals in a population have contact with each other. Infection levels can coincide with transport networks such as road and rail so individuals in areas with high levels of such transport links are more susceptible to catching infection. This randomness can exist at all levels, from random human
contact to random movement among cells of the immune system to even lower molecular levels where random variation exists due to Brownian motion of the interaction between molecules (Gaspard, 2005).

Capturing variation from computer simulation can be useful as this can provide a range of output values rather than a single value for the average. In many scenarios in the real world where resources are limited, simulations are often used to determine which resources to utilise and when to use them. Having this additional information can assist in making these decisions.

To conclude, traditional forms of analysis assume a regular world with defined boundaries and known cause and effect relationships. ABM is able to harness the local level interactions to gain insight into complex phenomena such as the transmission of infectious disease and the world of global economies [148].

6.2. Research Findings

Three areas of study are carried out by the experiments presented in this thesis:

- To understand the variation of ABM without any input parameter modification to SDM with Monte Carlo simulation.
- To compare the variation of ABM and SDM when input parameters are stepped across a defined range. This includes the effect of changing the population which is defined as a special parameter and is used to determine the effect of scaling for ABM.
- To compare the variation of ABM with real variation in data using historical data for influenza from the Centers of Disease Control.
The first experiment shows that the ABM Model captures natural variation in SIR without parameter modification. SDM is able to generate variation in the output results but is very sensitive to input parameter changes and in the case where all input parameters are varied, the output curve no longer matches the output curve for the original data and the variation in the output space is much larger compared to when individual parameters are varied. One special parameter changed in the first experiment is the population size. This is done across three different orders of magnitude to understand the scaling properties of ABM. It is found that it may be possible to use smaller population sizes combined with time changes in the environment for ABM to overcome the performance issues with running the simulations at full population level in order to complete the simulations within a feasible length of time.

For the second experiment where input parameters are varied for both models, SDM is able to produce a repeatable set of output results given any defined set of input parameters which can be useful as it makes the simulation reproducible. However due to the lack of variation there is no variation in the output space. As the input parameters are varied, there is a point in which the output curve no longer matches the empirical data. This is defined as the 'tipping point' in this thesis. There is one tipping point for when the input parameter is incremented and another for when the input parameter is decremented. The simulation from the SDM Model produces a single region for the tipping point. The simulation from the ABM Model is able to produce a range of output values for the upper and lower tipping points. These are analysed using box plots. From the experiments, it is found that the simulations from the ABM Model are able to produce upper and lower tipping points similar to those for the
simulations from the SDM model. These tipping points are maintained at different population sizes.

The third and last experiment compares variation in the output space of ABM to that of historical data. Real variation is highly complex and is dependent on a very wide range of factors, some of which are difficult to measure (such as the exact type of influenza) and many of which are difficult to measure (such as the contact network for the population).

It is found that the variation in ABM Model is comparable to that of historical data but this only applies when the same number of historical data sets is compared to the number of simulations. In this study, six years worth of historical data is compared to six simulations. When the number of simulations is increased to one thousand, the variation is no longer comparable as the averages for the output curve begins to smooth out. The experiment also found that the choice of parameters for the environment is important when generating variation that is similar to real variance.

6.3. Contribution to Knowledge

Knowledge of the range of peak infectivity in SIR is useful for managing and delivering resources in cases of outbreaks on infection. Standard SDM based models for SIR are simplistic and don't insight into the range of output values. Use of Monte Carlo simulation for SDM attempts to address the issue of variation outside of the inner workings of the SDM model.

Current research in the variation generated by ABM is limited. Although variance has been shown to occur [60], there is limited detailed analysis of the level of variation across time and limited analysis of how the variation compares to real variation.
present in real data. This thesis contains a detailed investigation of the variation generated by an ABM model in the context of the SIR model of infectious disease carried out through a series of experiments. This investigation can serve as useful set of practical information for any researchers planning to carry out similar. Key areas of research presented in this thesis cover:

- The ability of ABM Models to capture natural variation in infectious disease.
- An investigation into performance issues experienced with ABM Models and how these can be overcome by scaling down population sizes.
- An investigation into the side effects of scaling down population sizes for an ABM Model in terms of variation and total number of infections.
- The impact of environmental change on the output variation and total number of infected people over time.
- How variation in an ABM Model of SIR compares to real variation in SIR.

The experiments find that ABM is able to effectively capture natural variation of the SIR model of infectious disease without any changes to the input parameters. The variation generated by the model can be tailored and finely adjusted by modifying the environment of the model to meet the needs of the output variation if required without any changes to the main input parameters for the model.

The experiments with the three different population sizes show that it is possible to reduce the population size to make simulation time feasible using a combination of environment changes and the introduction of a time delay which can be used to align the output curves if required.
It is found that the variation generated by an ABM Model is affected by the number of simulations and the environment used. As the number of simulations is increased, the level of variation is reduced. Therefore, when attempting to match variation in a real system, the number of simulations used to monitor the variation is a factor that needs to be taken into account during validation of the model. It may be possible to use the same number of real world datasets and the same number of computer simulation runs in order to validate the initial ABM Model as performed in 5 of this thesis where the variation from an ABM Simulation of SIR is compared to real influenza data.

The environment has a significant impact on the variation. In the experiments, a small world network is used and it is found that the level of variation generated by the simulations is sensitive to the parameter values used for the small world network. Therefore, when generating the model, the type of network topology used and the parameters and initialisation for the network topology need to be taken into account during the validation of the output curve and the variation in the output results for the model. The choice of network topology depends on the real world system being modelled.

The small world network was chosen for the SIR experiments as it represented the most suitable network topology based on the type of connections in the population but other topologies including the use of custom network topologies. These can be implemented but in order to define custom networks, knowledge of the connections between people in the population is required and often this knowledge is difficult to obtain or identify, therefore a pre-defined network topology may represent a suitable abstraction for a real world system.
Although there have been a number of studies related to ABM and SDM including those by Rahmandad and Sterman [95] and also by Macal [60], these studies do not examine in detail the variance in the output from the simulations based on these methodologies. Instead, the variation presented in the studies as a useful by-product property of ABM and an aspect to consider when deciding which method to use. The research presented in this thesis takes a detailed look at the output from ABM and SDM with equivalent input parameters and when compared with real data. In the final chapter

The main focus of the work by Rahmandad and Sterman in the study of heterogeneity and network structure is around the decision of when to use ABM or SDM. This is done using a series of experiments examining the output curves of ABM across different network topologies and Differential Equations. The results for the ABM Models is presented as a series of output curves showing envelopes of 50%, 75% and 95% of 1000 simulations.

In contrast, in this thesis, the focus is on the variation generation by ABM and SDM with input parameter variation. The initial experiments presented in Chapter 3 are similar to the type of experiments carried out in the work by Rahmandad and Sterman where no parameter variation is applied. The work in the remaining two experimental chapters in this thesis is not carried out in the work by Rahmandad and Sterman. This is the work based on stepped parameter ranges and the comparison of the variation generated by ABM compared to real variation using real influenza data.

The study by Rahmandad and Sterman recognises the potential benefit of ABM which relaxes assumptions of aggregation but also recognises that use of ABM results in
computational costs which limits the amount of sensitivity analysis that can be performed. Additionally, the study recognises that ABM requires understanding the local behaviour of the agents. This can sometimes be difficult as the behaviour of the agents may not be well defined or easy to capture and also increases with the complexity of the model. As a result the study concludes that the choice of whether to use ABM or SDM is a trade-off between performance and the benefit of disaggregation. Detailed analysis of variation or the variation compared to real data is not carried out as it is outside the focus of the study.

Macal's study in 'To Agent-Based Simulation from SDM' examines how a model in SDM can be translated to ABM using formal specifications by identifying elements involving randomness which have been aggregated in the SDM Model and incorporating them in the ABM Model. The results are presented as output curves containing the output from SDM and the average values from ABM from 1000 simulations. There is an additional chart showing a 70% envelope around the output curves from ABM is presented. Detailed stepping experiments to determine variation or the comparison of the variation to real variation in empirical data is not performed as the focus is on the conversion process.

6.4. Limitations

The research carried out in this thesis is based on the SIR model. This is a model that generates a well known series of output curves for each of the values of infected, susceptible and recovered. In the experiments, the peak number of infections is used over time to carry out the experiments in variation. The shape for peak number of infections is a normal distribution curve so the analysis and conclusions discussed in
this work relate to a relatively simple type of output curve compared to some real world systems.

Real world systems can exhibit complex output curves. For example in the predator prey model, the output for the predator and the prey lead to oscillating curves for both the predator and the prey. Therefore, the level of variation exhibited by an ABM Model and the real system may vary. Other real world systems can exhibit chaotic output results. For example small changes in the random connections within the model can be amplified leading to results which may be unexpected from analysis of the local connections alone. One of the properties of ABM is that it is able to incorporate these low level interactions which can result in the emergence of unexpected output curves [149].

For the comparison of variation from an ABM Model compared to real data, historical influenza data is used over six years. Influenza is a highly complex infectious disease. As well as the occurrence of many different types of influenza and the nature of the infectivity which can occur as multiple waves across populations, it has some complex properties in that the virus for influenza is able to change its structure and form.

This makes monitoring and tracking difficult and brings into question the accuracy of the data which will always be an estimate. As the ABM and SDM models are validated against this data, there is the introduction of some level of uncertainty around the verification and validation of the data from the simulations and the real data. Real life system are always complex though and often it is difficult to trace all
aspects of the system. Therefore one way to overcome this is by collating larger sets of data. In doing this, the bias from small sample sets is reduced.

6.5. Further work

One of the properties of deterministic models such as SIR using SDM is that the model is total infection is able to become rebound due to the amplification caused by feedbacks in the system. This therefore makes them unrealistic for low values of infectivity [150] and is known as the fade-out phenomenon. In a more realistic model represented using agents, the number of infections can drop to none in which instance, there is no further infection unless there is a reintroduction of the infection from outside the boundary of the population being studied. Further work is necessary understand the impact of external factors on the output from an ABM Model. This can be useful for understanding scenarios such as the impact of infectivity based on cancellation of flights from areas of current infectivity. The outputs are heavily dependent on the network topology [151].

The choice of modelling paradigm to choose for a system usually depends on the purpose for which it is designed. Sometimes this can be a quick estimate based on well founded principles or confirmation of the workings of a system based on established principles. More often though, in an ever increasing complex world where data collection methods are improving to near real-time collection, there is often a need to try and understand the reasoning behind large complex systems as they occur. Questions such as:

- How did the infection start?
- How did the infection spread?
• How quickly did it spread?
• What was the contact profile of the infection?
• What is the current state of the infection for the population?
• What is the projected state of the infection for the population?
• What are the containment options?

Data collection combined with appropriate modelling is key to answering these types of questions. We need to be able to obtain information from the very start of the infection and this needs to be maintained in some central manner for the purpose of modelling. The data may be collected by health professionals during visits to clinics and this data can be fed into an ABM model based on a small world provide real time projections of infection transmission with the possibility to modify parameters and environments to determine the impact of changes in the environment and influenza infectivity on the peak infections across the population.

A number of studies have been carried out which make use of mobile devices to obtain infection information. Results show that data collected from mobile devices such as an iPhone running HealthMap compares well to data collected by Centres for Disease Control and Prevention [152].

Obtaining information for mobile applications in additional to traditional will provide a more comprehensive view of threats of emerging infectious diseases in a world becoming increasingly more interconnected [153].
References


[130] Halder, N., Kelso, J. K., & Milne, G. J. (2010). Developing guidelines for school closure interventions to be used during a future influenza pandemic. BMC Infectious Diseases, 10(1), 221.


influenza A (H1N1): a Monte Carlo likelihood analysis based on human mobility. BMC medicine, 7(1), 45.


