2020-2

A Hybrid Agent-Based and Equation Based Epidemiological Model for the Spread of Infectious Diseases

Elizabeth Hunter
Technological University Dublin, elizabeth.hunter@tudublin.ie

School of Computer Science

Technological University Dublin

A Hybrid Agent-Based and Equation Based Epidemiological Model for the Spread of Infectious Diseases

Submitted by:

Elizabeth Hunter

Supervisors:

Prof. John Kelleher

Dr. Brian Mac Namee

Thesis Submitted for the degree of

Doctor of Philosophy

February 2020

Abstract

Infectious disease models are essential in understanding how an outbreak might occur and how best to mitigate an outbreak. One of the most important factors in modelling a disease is choosing an appropriate model and determining the assumptions needed to create the model. The main research questions this thesis addresses are how do we create a model for the spread of infectious diseases that captures heterogeneous agents without using an inordinate amount of computing power and how can we use that model to plan for future infectious disease outbreaks.

We start our work by analysing and comparing equation based and agent-based models and determine that an agent-based model's stochasticity and ability to capture emerging results (complex and hard to explain results from interactions of agents) means that the agent-based model has an advantage in modelling the individual actions and complexities that make one infectious disease outbreak differ from another. Focusing on agent-based models, we take the model in two directions adding complexity and scaling up the model. Although adding complexity allows us to produce robust results, it increases run time so modelling anything beyond a small population is not feasible. Thus we focus on scaling up the model (from a town to a county) and determining what trade-offs need to be made to keep the model computationally tractable. With our scaled up model we look at characteristics of a town that come from its place in a network of towns, looking at how the centrality of a town affects how an outbreak spreads from a town and enters a town. We determine when a town has a high in degree centrality the

centrality of the other towns are not as important with respect to whether the outbreak will spread to the other towns.

The additional agents in the scaled up model lead to an extended run time. In order to reduce run time we make an assumption about the importance of heterogeneous mixing when there is a large number of agents infected and create a hybrid agent-based and equation based model that switches between an agent-based disease component and an equation based disease component based on a threshold of the number of agents infected. The hybrid model is able to save time compared to a fully agent-based model without losing a significant level of fidelity. This allows for the model to be scaled up to larger geographies and populations. Scaling the model to larger populations is essential in studying and testing the efficacy of interventions that would not be applicable at a smaller scale. To show this we use the hybrid model to analyse the effects of school closure policies across a network of towns, showing that closing both the town where an outbreak starts in and the town in the region with the highest in degree centrality can help mitigate an outbreak.

Declaration

I certify that this thesis which I now submit for examination for the award of PhD, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for graduate study by research of the Technological University Dublin and has not been submitted in whole or in part for another award in any other third level institution. The work reported on in this thesis conforms to the principles and requirements of the TU Dublin's guidelines for ethics in research.

TU Dublin has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature:

_____

Date:

_____

Acknowledgements

Most importantly, I would like to thank Professor John Kelleher and Dr. Brian Mac Namee for all of the support and guidance that they provided as supervisors throughout my PhD. This thesis would not have been possible without them.

I also would like to thank Technological University Dublin for providing me funding with the Fiosraigh Scholarship and the opportunity to take part in the PhD program. Thank you to everyone at TU Dublin who has played a part in my experience. I would like to express my gratitude to Dr. Flaminio Squazzoni and Dr. Pierpaolo Dondio for agreeing to be my examiners.

I want to thank my family especially my parents and my sister, Sarah, but also all of my extended family for their continued support and encouragement without who I would never have made it this far. They taught me that I could do whatever I put my mind to and have always instilled in my the importance of education.

I am so grateful to all of my friends both in Dublin and abroad that have supported me throughout my PhD. And thank you to the best house mate I could ask for, Dorina, and to Oscar for reminding me that sometimes you just need to take a break and go for a walk.

Table of Contents

1 Introduction **18**

1.1 Contributions 22

1.2 Thesis Summary and Structure 26

2 Epidemiology Models **28**

2.1 Epidemiology Concepts 28

2.1.1 Basic Reproductive Number 29

2.1.2 Effective Reproductive Number 30

2.1.3 Herd Immunity 30

2.2 Equation Based Models 31

2.3 Agent-Based Models 34

2.3.1 Modelling Disease 36

2.3.2 Modelling Society 42

2.3.3 Modelling Transportation 44

2.3.4 Modelling the Environment 47

2.3.5 Model Validation 48

2.4 Measles Dynamics 51
2.5 Conclusion 52

3 An Equation Based Model of Measles for an Irish Town 53

3.1 Data 54
3.2 Model 54
3.3 Model Evaluation 56
3.3.1 Modelling disease dynamics 57
3.3.2 Sensitivity Analysis 59
3.3.3 Case Study: Schull 2012 60
3.4 Simulating Additional Towns 62
3.5 Conclusion 65

4 Taxonomy of Agent-based Models for Infectious Disease Epidemiology 67

4.1 Disease Model 70
4.2 Society Model 71
4.3 Transportation Model 73
4.4 Environmental Model 74
4.5 Applying the Taxonomy 76
4.6 Conclusions 82

5 An Agent-Based Model of Measles in an Irish Town 86

5.1 Data 87

5.1.1 Population statistics 88
5.1.2 GIS data 89
5.1.3 School locations 90
5.1.4 Vaccination data 90
5.2 Model 91
5.2.1 Environment 92
5.2.2 Society 94
5.2.3 Disease 95
5.2.4 Transportation 97
5.2.5 Schedule 98
5.3 Model Evaluation 98
5.3.1 Modelling disease dynamics 99
5.3.2 Sensitivity analysis 100
5.3.3 Case study: Schull 2012 107
5.4 Simulating additional towns 112
5.4.1 Towns similar to Schull 118
5.5 Conclusion 124

6 Comparison of Agent-Based and Equation Based Models 127

6.1 Model Results 128
6.2 Experiment and Results 130
6.3 Conclusion 133

7 Methodology: Evaluating, Validating and Testing 136

7.1 Stochasticity and Confidence Intervals: Selection of Number of Runs 137

7.1.1 Experiment . . . 138

7.1.2 Conclusion . . . 141

7.2 Model Validation . . . 143

7.3 Model Testing . . . 145

7.4 Conclusion . . . 146

8 Adding Complexity to the Agent-Based Model 148

8.1 Segregation Models . . . 152

8.2 Assessing the Degree of Clustering by Socioeconomic Status in Irish Towns . . . 156

8.2.1 Residential Property Price Register data . . . 157

8.2.2 Calculating Dissimilarity in Irish Towns . . . 157

8.3 Modelling Irish Towns . . . 160

8.3.1 Calculating Dissimilarity In The Basic Town Model . . . 162

8.3.2 Using a Segregation Model to Better Model Dissimilarity . . 164

8.4 Results . . . 166

8.4.1 Using Segregation Modelling to Model Socioeconomic Clustering . . . 166

8.4.2 Assessing the Impact of Socioeconomic Clustering on Outbreak Modelling . . . 172

8.5 Conclusion . . . 175

9 Scaling up the Agent-Based Model **178**

9.1 Environment 179

9.2 Society 181

9.3 Transportation 184

9.4 Disease 186

9.5 Model Evaluation 187

9.5.1 Results 188

9.6 Experiments 189

9.6.1 Centrality 190

9.6.2 Town Similarities 193

9.7 Results 194

9.7.1 Degree Centrality 197

9.7.2 Closeness Centrality 201

9.7.3 Distance and Centrality 204

9.8 Conclusion 206

10 A Hybrid Model for the Spread of Measles **209**

10.1 Epidemiology Hybrid Models 210

10.2 Town Hybrid Model 213

10.2.1 Model Components 214

10.2.2 Model Evaluation 221

10.3 County Hybrid Model 233

10.3.1 Model Components 234

10.3.2 Model Evaluation 235
10.4 Conclusion 248

11 Testing an Intervention Strategy with the Hybrid Model 252

11.1 Agent-Based Models for Infectious Disease Interventions 253
11.2 Model 255
11.3 Experiments 255
11.4 Results 258
11.5 Conclusion 263

12 Conclusion 267

12.1 Summary 267
12.2 Research Questions 272
12.3 Future Work 276

Appendices 295

A Town Model Description 296

B County Model Description 315

C Hybrid Model Description 334

D List of Publications 357

E List of Employability and Discipline Specific Skills Training 359

List of Figures

2.1 The components of the disease model 38

3.1 Example SEIR curves. The curves are generated using a basic SEIR differential equation model. 58

3.2 SEIR curves from the Equation Based model. 58

4.1 Taxonomy of Epidemiological Agent-Based Models.
Grey branches and boxes outlined in grey are those combinations of component types that we did not find in our literature review and based on an analysis do not think would be feasible combinations. . 69

5.1 Example of Model Environment. The environment created by the model for the town of Schull, Ireland. The white lines are the boundaries of the small areas and the yellow agents are located at the agent households in the model. 93

5.2 Example SEIR curves. The curves are generated using a basic SEIR differential equation model. 101

5.3 SEIR curves from the model. SEIR infection curve for 10 runs of the model in a town in which no one was vaccinated or immune. . . 102

5.4 Outbreaks by probability of infection. Charts showing the change in average number of agents infected, the percent of runs leading to outbreaks and the maximum infected agents as the probability of infection changes. 104

5.5 Outbreaks by percent chance of staying home sick. Charts showing the change in average number of agents infected, the percent of runs leading to outbreaks and the maximum infected agents as the percent chance of staying home sick increases. 108

5.6 Average Number of Agents Infected in Schull by Age Groups. . . . 110

5.7 Distribution of Agents Infected by Week in Four Different Runs . . 111

5.8 Percent of Runs Leading to an Outbreak. Scatter plot of percent of runs resulting in outbreak and factors defining each town. *Outbreak* is percent of runs resulting in outbreaks, *small_areas* is the number of small areas in the town, *students* is the percent of students in the town, *unvaccinated* is the percent of unvaccinated agents in the town, *density* is the population density, and *transmission* is the probability of transmission per contact. 117

5.9 Histograms showing the percent of runs by number of agents infected for Schull, Shanagolden, and Strokestown 122

5.10 Schull, Ireland Map showing the population density per sqkm in Schull from the 2011 Census CSO (2014a). 123

5.11 Shanagolden, Ireland Map showing the population density per sqkm in Shanagolden from the 2011 Census CSO (2014a). 123

5.12 Strokestown, Ireland Map showing the population density per sqkm in Strokestown from the 2011 Census CSO (2014a). 124

7.1 Size of the Confidence Interval by Number of Runs for Kenmare, Tramore and Schull 140

7.2 Percent of Runs that Results in an Outbreak by the Number of Runs.141

8.1 Setup and Results of a Schelling type model in Netlogo. 154

8.2 The initial setup for the towns Schull and Tramore. 163

8.3 The setup for the towns of Schull and Tramore after the burn-in has been implemented. 168

8.4 Schull results for different levels of tolerance. The dashed lines are average initial dissimilarity index while the solid lines are average ending dissimilarity index. 170

8.5 Tramore results for different levels of tolerance. The dashed lines are average initial dissimilarity index while the solid lines are average ending dissimilarity index. 171

8.6 Distribution of total infected agents across the 300 model runs for Schull. 173

8.7 Distribution of total infected agents across the 300 model runs for Tramore. 174

9.1 Distance Matrix showing the normalized euclidean distance between towns 194

10.1 Distribution of the total number of time steps the disease model is equation based for the small area switch model. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based. 225

10.2 Distribution of the total number of time steps the disease model is equation based for the town switch model. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based. 226

10.3 Distribution of the total number of agents infected by run for the small area switch model 228

10.4 Distribution of the total number of agents infected by run for the town switch model 229

10.5 Distribution of the total number of time steps for the small area switch model to finish 231

10.6 Distribution of the total number of time steps for the town switch model to finish 232

10.7 Distribution of the total number of time steps the disease model is equation based when the model switches at the town level. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based. 237

10.8 Distribution of the total number of time steps the disease model is equation based when the model switches at the county level. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based. 238

10.9 Distribution of the maximum number of small areas that have an equation based disease component at a given time 239

10.10Distribution of the total number of infected agents when the county model switches at the town level. 241

10.11Distribution of the total number of infected agents when the county model switches at the county level. 243

10.12Distribution of the length of time taken for the model to finish when switching at the town level. 244

10.13Distribution of the length of time taken for the model that switches at the county level to finish 245

A.1 Image of Schull, Ireland. The white boarders are the boarders of the small areas that make up the town. 300

List of Tables

3.1 Vaccination Scenarios Sensitivity Analysis for the Equation Based Model 60

3.2 Population and model results for each of the 33 selected towns . . 64

4.1 Simulation Classification Table.

The *disease*, *society*, *transport*, and *environment* columns place the papers in our taxonomy while the *use* column details the intended uses of the model. The use of the model although not part of our classification system, can be affected by where a model falls in the taxonomy. 78

5.1 The 15 themes from the CSO census data tables CSO (2014b) . . . 88

5.2 Differences in model results based on changes in the probability of infection in the model. 104

5.3 Vaccination Scenarios Sensitivity Analysis 106

5.4 Differences in model results based on the percent chance of agents staying home when sick. 107

5.5 Area, population and other characteristics for each of the 33 selected towns . . . 116
5.6 Correlation table for percent outbreaks and the other town characteristics . . . 118
5.7 Percent outbreaks, area and population for each of the 12 selected towns and Schull . . . 119

6.1 Area, population and model results for each of the 33 selected towns 132
6.2 Correlation table for size of the outbreaks and the other town characteristics . . . 133

7.1 2011 population and town areas for the three towns . . . 138
7.2 Percent of runs leading to an outbreak and confidence intervals by town for 1,000 runs. . . . 138
7.3 Percent of runs leading to an outbreak and confidence intervals. . . . 139

8.1 House Prices from PSRA (2012) . . . 157
8.2 Dissimilarity index for each of the housing groups from PSRA (2012) 160
8.3 Starting dissimilarity index for each of the CSO social class groups . 162
8.4 Dissimilarity index for the Managerial and Technical CSO social class after segregation model burn-in process using different hyper-parameters. . . . 169

8.5 The starting and ending dissimilarity index for the all CSO social classes for Schull. The model used had a tolerance of 0.5 and a radius of 5. 169
8.6 The starting and ending dissimilarity index for the all CSO social classes for Tramore. The model used had a tolerance of 0.5 and a radius of 5. 170
8.7 Distribution of total infected agents across the 300 model runs for Schull 172
8.8 Distribution of total infected agents across the 300 model runs for Tramore 173

9.1 Percent of runs leading to an outbreak for the detailed environment model and the reduced fidelity environment model 181
9.2 Percent of runs leading to an outbreak for a steady versus a varying initial population 183
9.3 Summary statistic for the number of agents infected per run for a steady versus a varying initial population 184
9.4 Results for the town models and the county model 188
9.5 Normalized centrality by town 193
9.6 Percent of runs leading to an outbreak in each of the 16 towns when the outbreak starts in a random location or one of the eight selected towns. 195

9.7 Correlations between centrality and the range and average percentage outbreaks 196
9.8 Correlations between the percent of runs that lead to an outbreak and the total degree centrality of the town by starting location of the outbreak 197
9.9 Correlations between the percent of runs that lead to an outbreak in a town and the in degree centrality of the town by starting location of the outbreak 199
9.10 Correlations between the percent of runs that lead to an outbreak and the out degree centrality of the town by starting location of the outbreak 200
9.11 Correlations between the percent of runs that lead to an outbreak and the closeness centrality of the town by starting location of the outbreak 202
9.12 Approximate distances between each town in km 204

10.1 Percent of runs that lead to the model switching from agent-based to equation based for different versions of the hybrid model 224
10.2 Average number of milliseconds for a time step for different versions of the hybrid model switching at the small area level 227
10.3 P-values for the Wilcoxon rank sum test comparing the outbreak size distributions for the switching models to the completely agent-based model. 229

10.4 P-values for the Wilcoxon rank sum test comparing the outbreak time distributions for the switching models to the completely agent-based model. 232
10.5 Percent of runs that lead to the model switching from agent-based to equation based 236
10.6 Average number of seconds for a time step for different versions of the hybrid model 240
10.7 P-values for the Wilcoxon rank sum test comparing the outbreak size distributions for the switching models to the completely agent-based model. 243
10.8 P-values for the Wilcoxon rank sum test comparing the outbreak length distributions for the switching models to the completely agent-based model. 246
10.9 Percent of runs that the outbreak spreads to the given town 247

11.1 Towns in Leitrim with at least one school ordered by centrality . . 258
11.2 Towns in Leitrim with at least one school ordered by distance to Drumkeeran 258
11.3 The percent of runs that result in three or more agents becoming infected based off of the intervention strategies used in the model . 259
11.4 Summary statistics for the models with interventions, including the confidence interval for the mean 260

11.5 A comparison between the percent of runs that lead to an outbreak (2 or more infected agents) and the percent of runs where at least one agent is infected from outside of Drumkeeran the initial town. . 261
11.6 The percent of runs that result in three or more agents becoming infected when schools are closed by centrality. 262
11.7 The percent of runs that result in three or more agents becoming infected when schools are closed in towns by distance 263

A.1 State Variables for agents in the town model 299
A.2 State Variables for each grid cell in the town model 299

B.1 State Variables for agents in the county model 317
B.2 State Variables for each grid cell in the county model 318

C.1 State Variables for agents in the hybrid county model 337
C.2 State Variables for each grid cell in the hybrid county model 338

Chapter 1

Introduction

Epidemiology is defined as the study of the occurrence, distribution and determinants of health, disability and disease within a population. It plays an important role in public health allowing health professionals to gain a greater understanding of risk factors of both infectious and non-infectious diseases within a population. The knowledge gained from studying epidemiology can be used to implement control measures and identify important hazards that can be utilized in helping to mitigate the effects of a disease (Bartlett & Judge, 1997). Although epidemiology covers both infectious and non-infectious diseases this thesis focuses on infectious diseases.

The study of infectious diseases and how they spread is extremely important. The deadliest event in recent human history is not a war or a natural disaster, such as an earthquake or storm, but the Spanish flu outbreak that killed approximately 50 million people in 1918, 10 million more than those killed in World War 1 which

ended the same year (Taubenberger & Morens, 2006). It is commonly believed that the question concerning the next deadly pandemic is not if it will occur but when it will occur. The World Health Organization (WHO) prepares an annual review on priority diseases that they feel could be involved in a future public health emergency (WHO, 2018) and the UK's National Security Capability Review (NSCR) has added the threat from infectious diseases to the list of challenges that are expected to drive future security priorities (*National Security Capability Review*, 2018).

Large pandemics are not the only threat from infectious diseases, global travel and lower than desired vaccination rates means that outbreaks of diseases are still common in countries where the diseases are no longer endemic. With a more global population infectious diseases are more of a threat than ever. For example, within Europe, the first three months of 2018 saw 18,325 measles cases in 36 different countries with 23 deaths secondary to measles. While measles has been declared no longer endemic in Ireland, in the first 3 months of 2018 there were 59 measles cases reported in Ireland from three different outbreaks (O'Brien & Cotter, 2018)[1]. Further an outbreak of measles in Samoa starting in October 2019 has led to the country declaring a state of emergency with 1,644 cases including 20 deaths in a population of only 200,000 by the end of November with the outbreak still on going (Kelly, November 23, 2019).

One way to better prepare for an outbreak of an infectious disease is for epi-

[1]The World Health Organization (WHO) definition of a measles outbreak is two or more linked cases of measles (*Measeles: Vaccine-Preventable Diseases Surveillance Standards*, 2018)

demiologists and public health officials to learn as much as they can about a potential outbreak before it occurs by modelling potential trajectories of an outbreak in a population under different conditions. Modelling allows epidemiologists to run experiments on infectious disease spread without the moral, ethical and feasible problems that would come from running an experiment in the real world. In addition, as data is often not available, or incomplete due to under-reporting it is difficult to accurately estimate specific parameters, instead a range of parameters is often studied and models allow for experiments to be easily done for a variety of parameter values (Hethcote, 1989).

There are multiple methods that can be used to model the spread of infectious diseases. Two methods that are used most often and that we focus on are equation based and agent-based models. Both methods have advantages and disadvantages and can be useful in different scenarios, but one of the most important advantages of agent-based models are their ability to capture emerging results and interactions that equation based models do not (Hunter et al., 2018a). These interactions are important in capturing the many factors that influence the spread of an infectious disease. These factors include those related to the disease such as the the mode of transmission, the length of the latent or exposed period and the the length of the infectious period, and also include factors outside of the disease such as demographic, economic and geographic factors (Hethcote, 1989). However, equation based models are the model type that have been favoured historically and have been shown to both capture the macro level disease dynamics of the model and

are able to scale well to large populations.

As a model is only a simplified representation of a real world system in any type of model used it is necessary to make assumptions about what is necessary to include in the model. In many cases adding details to a model can provide robust model results but will lead to problems in terms of long model run times or complicated equation based models that are difficult to solve. In these cases a balance between the fidelity of the model and the computing power required to run the model must be found. There is no set threshold for this balance, but an analysis of the particular model and the changes in the results as the fidelity is altered needs to be done. Ideally there is a point where the fidelity is reduced but the model results are not drastically altered.

This thesis focuses on building large scale models of the spread of infectious disease that achieve this balance between high fidelity and feasible computational requirements. We use an agent-based model as the main portion of our infectious disease model because agent-based models are able to capture complex interactions between factors and emergent results based on agents' decisions within the model that other types of models cannot and we feel that these interactions and emergent results are essential in understanding the dynamics of an outbreak. However, when analysing the results and scaling up the model it becomes clear that even a simplified agent-based model can require a large amount of computing power. The main alternative models, equation based models, are not able to capture the emergent results and individual factors that an agent-based model does and

although less complicated without these emergent results too much fidelity is lost. In order to not lose as much fidelity but save time and computing power with equation based models we propose using a hybrid agent-based and equation based model. A hybrid model will allow us to create a model that still retains the advantages of an agent-based model and is scalable to larger populations while not significantly altering the results of the agent-based model. The main research questions we address are:

1 How can we create a model for the spread of an infectious disease for a specific population so that we accurately capture the heterogeneous characteristics of individuals so that it does not require an inordinate amount of computing power to run the model? This research question is answered in Chapters 5, 7, 9 and 10.

2 How can we use the model to plan for future infectious disease outbreaks? This question is answered in Chapters 9 and 11.

1.1 Contributions

The main contributions of this thesis fall into two main categories, contributions to the field of agent-based modelling research for infectious diseases and epidemiological contributions. The contributions to agent-based modelling research are:

- A taxonomy of agent-based models for human infectious disease epidemiology. The taxonomy is presented in Chapter 4 and published in Hunter et

al. (2017). When we began the research programme reported in this thesis there was no standard for creating an agent-based model to study the spread of an infectious disease through a population. Consequently, as one of the contributions to the field that this work makes, we developed a taxonomy of agent-based models for epidemiology based on a comprehensive review of existing models in the literature. That taxonomy can be used to both place an existing model within the wider body of agent-based models for epidemiology and guide in the creation of a new agent-based model.

- A methodology for validating and testing an agent-based model for the spread of infectious diseases is presented in Chapter 7. Standard methods for creating, validating, and testing agent-based models are not commonly found in the literature as the agent-based modelling field is so wide and so flexible. We outline a basic methodology that can be applied generally to agent-based models for the spread of infectious diseases. This methodology has been developed and applied through the work done to validate and test our own model. The methodology includes a process to determine the appropriate number of runs necessary to account for the stochasticity in an agent-based model using confidence intervals.

- A burn-in segregation model as an extra step in the agent-based model setup is presented in Chapter 8 and published in Hunter et al. (2018c). Often the data required to create a model might not be available or might not exist but if that data was included in the model it would effect the results.

We found that our data was not detailed enough to capture the clustering of agents by socioeconomic status but hypothesized that as socioeconomic status is related to vaccination rates clusters could lead to pockets of agents with lower than average vaccination rates or higher than average vaccination rates and these pockets could alter how the disease spreads. To overcome the problem we created an extra step in the setup of our agent-based model where an agent-based segregation model was used after the society was setup to allow agents to move homes and cluster with agents of similar socioeconomic status.

- A hybrid agent-based and equation based model architecture that combines the two existing hybrid modelling methods and is discussed in Chapter 10. There are few hybrid models in the literature and those that exist either completely switch between agent-based and equation based or have some component always agent-based and other components always equation based. Our model takes the approach that one component switches between equation based and agent-based at a certain threshold while the other components of the model remain agent-based. The approach allows us to keep the heterogeneous movements of agents while reducing the computing power required to run the model.

The following are the epidemiological contributions of the thesis:

- An agent-based model for the Irish context for the spread of measles. No other such model exists. Our fully agent-based model for Irish towns can be

found in Chapter 5 and published in Hunter et al. (2018b), our fully agent-based model for Irish counties can be found in Chapter 9 and our hybrid agent-based and equation based model can be found in Chapter 10. As there are often factors unique to a given population that play a role in how susceptible that population is to an outbreak or how well a given intervention is going to work for that population, it is important to have a model that is specific to the society that is being studied. Our model reproduces the Irish population using Irish census data and accurately represents towns with the correct population, number of schools and distances to other towns within the county.

- With our model we do an analysis of how the centrality of a town within a network of other towns affects the spread of an infectious disease through the network in Chapter 9. The work is published in Hunter et al. (2019). We look at a number of types of centrality and determine some relationships between the centrality and how the infectious disease will spread. The results of our centrality study can be used to influence decisions on how to slow the spread of an outbreak.

- School closure policies are used as a response to outbreaks, however, their use is often debated. While analysis of school closure policies exists in the literature, including analyses with agent-based models, we take a different approach in Chapter 11. Instead of looking at closures in a single town, where the outbreak starts, we use the results from our centrality analysis to

take a new approach to school closures and look at closing schools in multiple towns. Our results show that school closures can help to reduce the spread of an infectious disease but that there are multiple interconnected factors that need to be considered when creating such a policy.

1.2 Thesis Summary and Structure

This document is organized into twelve chapters, including this chapter, and three appendices. Chapter 2 provides background on epidemiology models, measles dynamics and reviews the agent-based modelling literature. In Chapter 3 we present our equation based model for measles spread. Chapters 4 and 5 focus on agent-based models with Chapter 4 presenting a taxonomy for agent-based models for human infectious diseases that is designed to help place an existing agent-based model into the existing literature and also to aid in the creation of an agent-based model. Chapter 5 presents our agent-based model for the spread of measles in a town. Chapter 6 discusses the comparison of the results from the equation based model and the agent-based model looking at the advantages and disadvantages of both.

As the advantages of the agent-based model seem to outweigh the disadvantages, Chapters 7 through 9 focus on our agent-based model. In Chapter 7 we present a methodology for model evaluation. As part of this methodology we determine the number of model runs needed to account for the stochasticity in the model. Chapter 8 deals with adding extra complexity to the town agent based

model and Chapter 9 discusses the assumptions necessary to scale up the model to the county level. In Chapter 10 we introduce a hybrid model that combines the characteristics of both the equation based model and the agent-based model and in Chapter 11 we use the hybrid model to test intervention strategies designed to lessen an outbreak. Finally, Chapter 12 summarizes our finding and future directions of the work. The three appendices provide detailed model descriptions for the town model, the county model, and the hybrid model.

Chapter 2

Epidemiology Models

Although there are many kinds of epidemiology models in this thesis we focus on infectious disease models and in particular our models are for the spread of measles. The following sections give a brief overview of epidemiology modelling. Section 2.1 gives a general overview of epidemiology modelling concepts, Sections 2.2 and 2.3 discuss in greater detail the two main methods for modelling infectious disease outbreaks: equation based and agent-based models, and Section 2.4 explains measles disease dynamics.

2.1 Epidemiology Concepts

In this section we describe a number of epidemiology concepts in detail that are related to modelling infectious diseases and are used throughout the thesis.

2.1.1 Basic Reproductive Number

R_0 is the basic reproductive number and is defined as the expected number of individuals infected by one infectious individual in a completely susceptible population. It is the standard measure of transmission potential of a disease (Thomas & Weber, 2001). Other factors such as the transmission rate or the infectivity rate of a disease are less frequently estimated than R_0. R_0 is typically calculated for outbreaks as it is thought of as one of the most important parameters when trying to control an outbreak (Ridenhour et al., 2014). Data from real outbreaks can be use to estimate R_0 (M. Keeling & Grenfell, 2000). However, the infectivity rate requires more information and detailed contact tracing to determine the proportion of susceptible contacts an infectious individual would infect (Ridenhour et al., 2014). Therefore, R_0 is the most commonly found parameter for disease dynamics and what we use in our modelling work. The parameter can be broken down into three components: number of contacts per unit time (c), the transmission probability per contact (τ), and the duration of the infectiousness (d). The relationship can be seen in Equation 2.1 (Thomas & Weber, 2001).

$$R_0 = c\tau d \tag{2.1}$$

The formula for R_0 is sometimes written in terms of other variables commonly used in equation based models: the rate of transmission per unit time (β) and the mean infectious period ($\frac{1}{\gamma}$).

$$R_0 = \frac{\beta}{\gamma} \tag{2.2}$$

β can be equated to $c\tau$ and $\frac{1}{\gamma}$ can be equated to d, thus Equations 2.1 and 2.2 are the same just written in terms of different variables. The formula that is used is typically determined by what variables are available.

2.1.2 Effective Reproductive Number

The effective reproductive number, R_e, is the reproductive number, R_0, adjusted to account for immunity in the population. It can be calculated using the following formula, where x is the percent of susceptible individuals in the population:

$$R_e = R_0 * x \tag{2.3}$$

2.1.3 Herd Immunity

Herd immunity is the concept that there is a critical number of individuals that need to be vaccinated or immunized to interrupt the transmission of an infectious disease in a population. The formula to determine the proportion of individuals who need to have been vaccinated to achieve herd immunity is (where R_0 is the basic reproduction number we introduced in Section 2.1):

$$p = (1 - \frac{1}{R_0}) \tag{2.4}$$

However, as the vaccinations are not always successful, the proportion needs to be adjusted by the vaccine effectiveness in order to find the critical vaccination coverage needed to reach herd immunity. The equation for the critical vaccination coverage is (where V_e is the vaccine effectiveness rate, and V_c is the critical vaccination coverage or the proportion of individuals who need to have been vaccinated to achieve herd immunity when we take vaccination effectiveness into account):

$$V_c = \frac{(1 - \frac{1}{R_0})}{V_e} \tag{2.5}$$

2.2 Equation Based Models

Historically equation based models have been used to model the spread of infectious diseases. The most common type of equation based model used for infectious disease modelling is the compartmental model, which is made up of a set of differential equations (Hethcote, 2000). The population in a compartmental model is assumed to be homogeneous, well mixed, and split into compartments based off of health status. Each compartment is defined with its own differential equation (Duan et al., 2015). The simplest compartmental model is the SIR model where the population is split into three compartments: susceptible individuals (S), infected individuals (I), and recovered individuals (R). The following equations define the system:

$$\frac{dS}{dt} = \frac{-\beta SI}{N} \tag{2.6}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{2.7}$$

$$\frac{dR}{dt} = \gamma I \tag{2.8}$$

where N is the population size, β is the rate of transmission per unit time, γ is the recovery rate, and $S(t) + I(t) + R(t) = N$ (Hethcote, 2000).

Typical variations of the SIR model include the SEIR model (susceptible, exposed, infected and recovered), the SIS model (susceptible, infected and susceptible) and the SIRS model (susceptible, infected, recovered and susceptible). The models can be made more complicated and realistic by adding additional compartments for various characteristics of agents including age groups or vaccination status. These models can be used to better understand the dynamics of a disease. Hogan et al. (2016) create an age structured model for Respiratory Syncytial Virus, a common childhood infection, where each age group has its own compartments. The model can be useful when simulating age-dependent interventions such as vaccination. The effects that vaccination rates have on measles outbreaks are studied using the Pang et al. (2014) model.

While the majority of equation based models in the literature utilize differential equations another possible form of an equation based model that can be used to model the spread of infectious diseases is a difference equation or discrete time model. They are a type of mathematical model that is similar to a differential equation but are over a discrete time space instead of a continuous time space, as is the case with differential equations. Difference equations can exhibit behaviour

that a differential equation can not, with even a simple non-linear difference equation being able to show chaotic behaviour (Allen, 1994). While the majority of epidemic models in the literature are differential equations, there are some that use difference equations to better capture the dynamics of an outbreak. An example of a basic SIR difference equation model is as follows:

$$S_{t+1} = S_t - \frac{\beta I_t S_t}{N} \tag{2.9}$$

$$I_{t+1} = I_t + \frac{\beta I_t S_t}{N} - \gamma I_t \tag{2.10}$$

$$R_{t+1} = R_t + \gamma I_t \tag{2.11}$$

where N is the population size, β is the rate of transmission per unit time, γ is the recovery rate, t is the current time step, $t+1$ is the next time and $S_t + I_t + R_t = N$.

The epidemiology literature contains many reports of equation based models being used to analyse a specific outbreak or epidemic after the fact. These models are often used to determine lessons learned from the outbreak. For example, Vaidya et al. (2015) model the spread of H1N1 in a rural university town and determine that a portion of the susceptible population was protected from infection through self-isolation, social distancing or other preventative measures and this protected population played a substantial role in the dynamics of the epidemic. Ketema et al. (2015) show that to best capture the dynamics of Ebola spreading in West

Africa an additional compartment, isolated, is needed since isolation is commonly used in Ebola cases. Equation based models have also been used to help shape policy during an outbreak. A series of models were used to help inform policy decisions to control the 2001 foot-and-mouth disease epidemic in the UK (Kao, 2002).

Although equation based models have proven to capture the macro level dynamics of an infectious disease outbreak and have been used in the development of control policies and responses to outbreaks, there are some disadvantages to using an equation based model. Equation based models can not provide detailed information on the spread of the disease. In addition, the small set of variables that are used in an equation based model may not be enough to define an outbreak. Assuming that the population is homogeneous within a compartment can also be a problem in not capturing the individual variations and actions that can have a major impact on the course of an outbreak (Duan et al., 2015).

2.3 Agent-Based Models

One main alternative to modelling with equation based models is to model an infectious disease outbreak with agent-based models. Agent-based models are an important tool in studying the dynamics of infectious diseases, they allow the user to capture emerging results and interactions that might not otherwise be captured in an equation based model. Agent-based simulations are already being used to help decide on policy in the models by Barrett et al. (2008), Aleman et al. (2011)

and Lee et al. (2008). Other models are being used to understand past outbreaks so as to be better prepared in the future, such as Merler et al. (2015)'s model of the Ebola outbreak in Liberia and Friás-Martínez et al. (2011)'s model of the H1N1 outbreak in Mexico City.

As new infectious diseases emerge, agent-based models can be used as an aid to help understand how a population can be affected and how we should react to an outbreak. To do this it is necessary to have a strong understanding of all possible factors in disease spread. Much of the research being done now with agent-based models helps us to get to that point. For example, Epstein et al. (2008)'s work on fear leading to agents fleeing the area of an outbreak and spreading the disease further could play an important role in future simulations of emerging diseases.

As there is no obvious standard in creating an agent-based model to study infectious disease spread it is necessary to do a comprehensive study of the existing models in the literature in order to understand where the gaps in the literature are and where any model created will sit in the existing state of the art. The following sections were first presented in Hunter et al. (2017). From that study we have determined that there are four main components of an epidemiological agent-based model: disease, society, transportation, and the environment. Although we separate them for the purpose of understanding the agent-based epidemiological model, the components are intertwined. In reviewing the literature we focused on how different models treat these four main components, starting with the disease component in Section 2.3.1, then the society component in Section 2.3.2, the

transportation component in Section 2.3.3, and the environment in Section 2.3.4 and finally we look at how the papers present model validation in Section 2.3.5.

2.3.1 Modelling Disease

The agent-based modelling literature tends to treat infectious diseases in one of two different ways. Research is either done to create a general model where the disease parameters can be changed to show how various diseases will spread through different populations or the research focuses on modelling a specific disease and often a specific outbreak of that disease. A general disease model should be adaptable to multiple diseases of the same form of transmission, typically airborne transmission. These general disease models make sense in a scenario where the modellers want to create a tool to study future potential outbreaks. This way the model can be adjusted to different disease dynamics based on what disease is to be studied without creating a new model each time. For example, Barrett et al.'s (2008) Episimdemics was created for the US without a specific disease so that it could be adjusted for different possible outbreaks. It has been used to help determine policy in the face of a pandemic in the US.

A specific disease model allows for a model to better capture specific disease dynamics. While general disease models typically stick to airborne transmission, specific models can take into account other transmission methods such as water borne infections. Specific disease models can also take into account factors that might influence the spread of a given disease such as including infection during

funerals in the model of Ebola spread in Liberia by Merler et al. (2015).

Regardless of the specific or general nature of the model, the disease model will have many of the same components. The breakdown of the components can be seen in Figure 2.1. Disease models for agent-based models are broken up into two parts: between host transmission and within host progression. Between host transmission occurs when a susceptible agent comes into contact with an infectious agent, and the between host transmission component of a disease model simulates how a disease is transferred when this occurs. The within host progression component of a disease model simulates how, when an agent becomes infected, they move between the different states of the infection (for example exposed, infected, and recovered). Both parts of the disease model are important in accurately simulating how a disease will spread.

The transmission dynamics are a key factor in how the disease spreads between individuals. Disease can be spread through human-to-human contact, to humans from food or drinking water, or between hosts of different species, for example mosquitoes to humans. When an agent comes into contact with an infected agent, infected food or drinking water or an infected species, a probability distribution is used to determine transmission. The transmission can be affected by a number of factors outlined in Figure 2.1: transmission dynamics, society, transportation and environment, and behaviours. A number of agent-based models dealing with specific diseases contain different transmission dynamics based on the disease being modelled. Some of these models are for diseases such as cholera or malaria that are

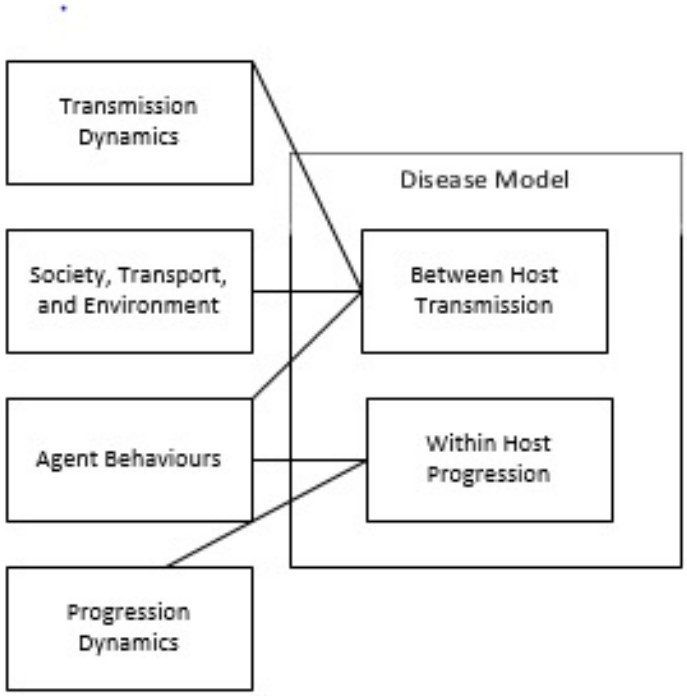


Figure 2.1: The components of the disease model

not spread through human contact but through contaminated food and drinking water or through insect bites. In the model by Crooks & Hailegiorgis (2014) for the spread of cholera in a refugee camp, agents excrete a certain amount of the cholera bacteria based on the stage of infection that they are in. Water contamination is determined through a hydrology model where the flow of rain water affects the total amount of pollutant and if an agent drinks contaminated water they have a certain probability of becoming infected. Malaria is spread by mosquitoes thus any simulation for malaria would have to take into account not only human movements but also mosquito movements (Linard et al., 2009). In addition, models can alter different transmission dynamics by agents for example by making some agents super spreaders, individuals who are more infectious than other individuals (Duan

et al., 2013).

The parameters of the society being modelled have a major effect on how a disease will spread between hosts. A densely populated area will result in more contacts between agents, and thus a greater likelihood of infection (Perez & Dragicevic, 2009). The social networks of agents also influence the disease spread. In their model of a disease spreading through a small Australian town, Skvortsov et al. (2007) found that the majority of infections in their model occurred at the schools. This was because every agent at a school was in contact with every other agent at the school. The large social networks of agents in the model led to a higher infection rate.

Agents' behaviours can also have an effect on the between host progression of a disease. For example, if the agents respond to an outbreak or possible outbreak by fleeing they may be spreading the disease at a greater rate than if they stayed home in isolation (Epstein et al., 2008). Alternatively, if agents choose to isolate themselves once infected, they reduce the number of contacts they make and thus reduce the number of agents the disease spreads to (Dunham, 2005). Diffusion of information about a disease can lead to agent's taking part in preventative behaviours such as getting vaccinated or taking medicine, such as flu prophylaxis, that will reduce the chance of infection (Mao, 2014). When modelling the Ebola epidemic in Liberia, Merler et al. (2015) included change in agents behaviours based on information about the disease. In the real epidemic as people learned that Ebola spread at funerals and to health care workers and other non-Ebola

patients in hospitals the number of safe funerals increased and health centres that only treated Ebola patients opened thus reducing transmission of the virus. This was reflected in the model with the number of hospital beds changing as time went on and the probability of becoming infected at funerals decreasing (Merler et al., 2015).

Within host progression does not have as many outside influences as between host transmission. The make-up of the society has no effect on how an agent moves from exposed to infected or infected to recovered. Similarly behaviours of other agents have no effect on within host progression: while an infected agent deciding to stay home from school or work might reduce the chances of other agents becoming infected, other agents' actions will have no effect on the progression of a disease within an agent (Mao, 2014). On a basic level all of the within host progression models are similar. A disease will move between states based on a probability distribution. Many of the agent-based models use a form of the SIR model, introduced in Section 2.2, to simulate disease progression (Dunham, 2005; Friás-Martínez et al., 2011; Merler et al., 2015; Perez & Dragicevic, 2009; Rakowski et al., 2010a; Crooks & Hailegiorgis, 2014). The SIR disease progression is based off of the different components in the SIR equation based models. The SIR model categorizes individuals into susceptible, infected or recovered states and looks at movement of individuals between these. Variations of the SIR model can include additional stages such as exposed (M. J. Keeling & Rohani, 2008). Although the SIR form of the disease model can be used for a specific disease it is often used when

a simulation is created for a general disease. A few models take more complicated disease dynamics into account moving away from the basic SIR type model. In modelling tuberculosis (TB), Tian et al. (2013) include states particular to TB including high and low risk latently infected agents, latently infected with previous treatment agents, undiagnosed infectious and non-infectious agents, active TB agents, and active undiagnosed infectious and non-infectious TB with previous treatment agents.

Although other agents do not have an effect on within host progression, the infected agent's behaviours can have an effect on the progression. Preventative behaviours can reduce the chance of an agent moving between susceptible and infected or increase the chance of an agent moving between infected and recovered (Mao, 2014). In some models having been vaccinated can reduce the chance that an agent moves from exposed or latently infected in the case of tuberculosis (Tian et al., 2013).

The factors affecting the transmission or progression can vary between the models but typically fall into the categories of progression dynamics, behaviours and society factors. However, the general disease models will have simple transmission and progression models while specific models tend to have more complicated transmission and progression models that reflect the given disease. The more factors added into the model the more realistic it will be. This, however, comes at a price and increased model complication leads to increased computational resource requirements. Thus it may become necessary to make trade-offs especially for larger

scale models between detail and computing power.

2.3.2 Modelling Society

In the spread of an infectious disease one of the main components that can have an effect on the course of the outbreak is the structure of the society. The number of people or agents, the household structure, number of students in each school, number of schools and workplaces are all things that need to be considered when simulating a society. It must be determined if the model will simulate an existing society or if it will be more general. We consider any simulation that uses real data to model a society a specific society model and any model that generates a society without the use of real data a general society model. General society models can be made by randomly placing agents in an environment. For example, the model by Dunham (2005), was created by generating 50 genderless and ageless agents and having them commute back and forth from their home locations to their work locations. Similarly Perez & Dragicevic (2009) created a society by randomly assigning genderless and ageless agents to a residential area and then randomly dividing that population into workers and students. The advantage of creating a general society model is that it does not require the large amounts of data necessary to simulate a real society. Because the data is not needed it will take less time for a modeller to begin the process of creating the simulation, the initialization of the simulation will need less computer power and time and it will be easier to scale up the model to a larger population.

In order to create a simulated society, model based on a real society simulation designers typically take census data from the population they are planning on recreating or from a similar population. For example, Skvortsov et al. (2007) used census data to determine the age/sex breakdown of the actual population of an Australian town and had the model build families based on average family size obtained from the census data. To model the spread of influenza through Poland, Rakowski et al. (2010a) use census data to assign individuals to a family based on age and relationships: a child will only be assigned to a house if an adult is already living there and the probability that two adults will live in the same house depends on the attraction which is determined by the difference in age between them. The scale of a specific society simulation can range from a small town (Skvortsov et al., 2007), to a community (Lee et al., 2008), to a region (Aleman et al., 2011) or to a country (Ajelli et al., 2010). As it is important to capture the social networks of an individual to determine the path of a disease spreading through a society, some models can differentiate between close contacts (other agents at home or work) and occasional contacts (agents in service places such as shops) (Crooks & Hailegiorgis, 2014). The social networks of agents in the society can also be broken down into weekday and weekend networks as it is more likely that an agent will interact with co-workers during the week and with friends during the weekend (Friás-Martínez et al., 2011).

Specific societies have a more obvious interpretation and use: their results can be applied to a given population to help make decisions about future outbreaks

or learn from past outbreaks. However, in order to create such a simulation data is needed and the more realistic a simulation is the more detailed data that is required. General society simulations may not require any data at all and thus can be easier to create in situations where data is scarce or hard to access.

The scale of the model must also be considered: for a simulation of a specific society the scale (country, region, city etc.) will be determined by the society being simulated. However, for a general model it is necessary to determine how many agents will be used in the simulation. The scale chosen for the society can also have an effect: the larger the scale the more computing time the simulation will take to run. However, small scale societies, particularly small scale general society simulations, may not have as much realistic interpretability as it would be difficult to find a real world application for such a model. The way that society is simulated will influence the rest of the model including how transportation is simulated and how the results of the model will be interpreted and used.

2.3.3 Modelling Transportation

The majority of agent-based epidemiological simulation models contain some form of movement or transportation of agents through the model environment, and choices must be made about how to simulate this. The majority of simulations drive agent movements based on the society model and the agent behaviour rules. Typically an agent will simply move from the house to which they are assigned to their workplace every day. Some more sophisticated models, however, also

include destinations such as markets, shopping malls, pubs, friends' homes, health centres and religious centres (Crooks & Hailegiorgis, 2014; Mao, 2014; Perez & Dragicevic, 2009; Simoes, 2006) and simulate agents movements between these locations following a weekly schedule.

The transport model in a simulation governs the way in which agents move between different locations. Simulations can use a very simple transportation model where agents simply move between locations in a straight line at a constant speed (Dunham, 2005). More realistic transport models use geographic data containing information about transport infrastructure to plan routes between destinations following footpaths and roads. Some models require the agents to select the shortest route while others allow less optimal travel (Crooks & Hailegiorgis, 2014; Perez & Dragicevic, 2009). It is possible to have more specific data to model movements such as cell phone data where an individual's real movements can be tracked based on where a phone call or other telephone service is used (Friás-Martínez et al., 2011). However, this kind of data is not easily accessible to all researchers. Some of the most sophisticated transport models include public transportation, as public transportation can be a crowded location where diseases are transmitted (Rakowski et al., 2010a; Aleman et al., 2011).

Movement can be affected by the agent's choices and behaviour. For example, if an agent is infected a model can allow the agent to decide if they are going to take a sick day (Dunham, 2005). A model by Crooks & Hailegiorgis (2014) went further allowing agents to set goals based on an agent's attributes and needs

that determine movement. Travelling longer distances can also be considered. A model for the spread of mumps in Portugal (Simoes, 2006) not only considers neighbourhood and intra-region travel, but also has a component for inter-region travel. In modelling influenza epidemics in Poland, Rakowski et al. (2010a) assign a certain number of agents at each time step to traveller status. These agents then choose their end points, transfer cities, and co-travellers. Co-travellers and the number of co-travellers are chosen randomly to simulate both public and private transportation. The movements of agents can have a great effect on the outcome of a simulation. Movements will determine who an agent contacts and thus affect how a disease will spread. Some of the advantages of including transport in the model is the ability to capture the location of infections. This could help identify potential 'hotbeds' of infection such as schools. It also allows for more realistic interactions outside of an agent's family or friends network.

There are some infectious disease models that do not include transportation of any kind. These models rely on contact networks to determine the spread of the disease. Agents who are in networks with other agents have a probability of coming into contact and spreading the disease. Tian et al. (2013) use such a model for their TB analysis and Olsen & Jepsen (2010) similarly create a model without transportation to model the spread of HPV. If the disease dynamics are not as reliant on agents' day to day movements then not including transportation in the model will lead to a faster run time for the model.

2.3.4 Modelling the Environment

The environment is an essential part of the agent-based model as it is where the agents move and interact. However, the level of complexity of the environment can vary based on the needs of the model and the transmission dynamics of the disease. An environment model can be as simple as a spatial grid upon which agents are placed as in Duan et al. (2013) and Dunham (2005). Simple environment models are easy to set up and run, however, they give little more information on the contact patterns of agents then you would get from an equation based model while models with added environment are more capable of capturing heterogeneous mixing. Slightly more complicated environments can include buildings for workplaces and schools, roads, and residential areas. Lee et al. (2008) creates an environment that includes schools and residential areas. These models have an advantage over the grid environment model as realistic movement patterns can be simulated and high infection areas can be determined. Lee et al. (2008) use their model to determine when school closing policies should go into effect and how long the closings should last. Such analysis could not be done with a simple grid environment: the inclusion of the schools and workplaces allows the modellers to accurately capture where agents become infected. Along with the buildings and roads many models also use other GIS data including elevation data to create their environment such as the models by Ajelli et al. (2010), Barrett et al. (2008) Crooks & Hailegiorgis (2014) and Simoes (2006). This allows for an accurate representation of the town, city, country etc. that is being modelled.

However, some models require a more detailed environment beyond roads and buildings as the environment can have an impact on disease transmission. Sophisticated environment models can also include factors such as temperature or precipitation or other populations that help to spread a disease can be included in the model. Linard et al. (2009) not only include mosquitoes in their malaria model but also temperature, water levels and vegetation levels. In order to model the spread of cholera through a refugee camp, Crooks & Hailegiorgis (2014) include a hydrology model as cholera is a disease spread through the consumption of infected water. Adding the additional environmental factors is essential to model some diseases, such as cholera and malaria, but the transmission dynamics of other infectious diseases can also be influenced by environmental factors. For example, influenza outbreaks most often occur in the winter months. Including environmental factors in an agent based model may capture factors in disease transmission that may have otherwise been ignored. However, the more factors that are included in the model the more complicated it becomes. As agent-based models tend to be computationally intensive additional factors can lead to difficulties in running the model.

2.3.5 Model Validation

One of the most important issues for agent-based modelling is validation. If a model is not validated, any surprising results cannot be completely trusted. There is currently no exact definition or methodology to test the validity of an agent-

based model (Richiardi et al., 2006). For epidemiological models, it is possible to simulate an infectious disease outbreak that has occurred in the past. In these cases validation can be possible through comparing the simulated outbreak with the real outbreak (Olsen & Jepsen, 2010; Merler et al., 2015; Crooks & Hailegiorgis, 2014; Perez & Dragicevic, 2009). This gives confidence that the model correctly simulates the dynamics of the disease and the society, allowing the researcher to make the assumption that the model will simulate future outbreaks correctly providing insight into the disease behaviour. For a common infectious disease such as influenza, data sources such as Google flu trend statistics can be used to supplement lab-confirmed reported cases in validation as the Google data will pick up some cases that are not reported (Mao, 2014). It is possible to use other statistics besides prevalence to validate an agent-based model. For example, if real movement data is available a comparison can be made between the real movements of individuals and the movements of agents (Friás-Martínez et al., 2011).

However, if the model is not simulating a past outbreak or epidemic (which is often the case with a general disease model) there are ethical, logical, and practical constraints to getting data for validation: it is not feasible to run an experiment to determine how an infectious disease will spread through a population (Hernán, 2014). One alternative to using real data for validation is cross validation: the output of an agent-based model can be compared to the output of another widely used model such as an equation-based SIR model. The number of susceptible, infected and recovered individuals over the simulated period can be compared between the

two models. Although it is likely that the numbers will not match exactly due to differences in model assumptions, if the infection curves, representing the number of susceptible, infected and recovered individuals at each time step, follow a similar trajectory it is likely that there is some validity in the agent-based model. If a simple agent-based model is validated with equation-based models it is possible for the researcher to add additional factors into the agent-based model. The results of the expanded model can be analysed knowing that the basic disease dynamics of the model were validated (Skvortsov et al., 2007; Rakowski et al., 2010a). One other alternatives to validation is determining adequacy (Apolloni et al., 2009). Adequacy is the idea that the appropriate and informed decisions are made when creating the model. When considering adequacy it is important to question if any new input in the model decreases uncertainty by improving precision of the final results and if the input significantly changes the model (Xia et al., 2013).

Although validation is an important step in creating any model, authors do not always include the validation process in their papers. For example, Barrett et al. (2008) mention that the model has been validated but do not describe the validation process. Some papers such as Dunham (2005) do not refer to validation at all. While other papers only briefly mention the comparison to real data in the results or discussion section of the article such as in Crooks & Hailegiorgis (2014) and Perez & Dragicevic (2009). In this thesis we will approach validation of our models in a number of ways. In Chapters 3, 5 and 10 we cross validate our models against a simpler model. Additionally in Chapters 3 and 5 the results

of our models are compared to real outbreak data as a source of validation. We also use the concept of adequacy to help validate our model in Chapters 3 and 5 looking at how changes in parameter values such as vaccination rates change the output of the model. In Chapter 7 we use the process and methods that we have gone through to validate our models to present a general methodology for validating agent-based models.

2.4 Measles Dynamics

In this work we study measles-like diseases. Measles is a highly infectious virus that is often associated with childhood. It is most common among children between the ages of 1 and 4 years old but anyone who is not immune from previous illness or vaccination is susceptible (HSE, 2017). Once a person is exposed to measles the disease has a 10-14 day incubation period before the onset of the first symptoms. Fourteen days after initial exposure, the measles rash occurs (Nelson & Williams, 2007). An individual can be infectious between two and four days before the onset of the rash and five days after the onset of the rash (HSE, 2017). R_0 has been calculated for measles in a number of studies using real world measles outbreak data and has been found to be between 12 and 18 (Nelson & Williams, 2007).

2.5 Conclusion

This chapter has introduced a number of fundamental concepts that will underpin many of the later discussions in this thesis, these include: the general concepts of epidemiology models that we have discussed in Section 2.1, the dynamics of the disease being modelled detailed in Section 2.4 and the state of the art for both equation based and agent-based models which we reviewed in Sections 2.2 and 2.3. We will use the epidemiology concepts and the measles dynamics in creating our models in Chapters 3, 5, 9 and 10. The literature reviews aided us in placing our model in the state the art and, as we will discuss later, informed the development of the taxonomy for agent based models that is presented in Chapter 4. The taxonomy uses the four main components of an agent-based model for infectious diseases discussed in Section 2.3 to classify existing agents based models and can also help to serve as a guide in the creation of a new model. We will use it in the development of our model in later chapters.

Chapter 3

An Equation Based Model of Measles for an Irish Town

As equation based models have been used historically to model outbreaks and have been shown to be able to capture the dynamics of an infectious disease outbreak in this thesis we use an equation based model as an initial baseline for our research. This chapter reports an experiment that used an equation based model to model the spread of an infectious disease, specifically measles. In the following sections we present a compartmental differential equation SEIR model for 29 different age groups that includes vaccination for the spread of measles. Differential equations are used over difference equation for this study as it is the more popular and well used equation based modelling method used to study infectious disease outbreaks. The model is tested on a number of small to medium size Irish towns. The following section describe the data used to create the model. Then we describe the model,

evaluate the model using sensitivity analysis and a case study and finally show results simulating 33 different towns in Ireland. This model and the following results were first presented in Hunter et al. (2018a).

3.1 Data

Although equation based models are not data intensive, a number of data sets provide information for the starting conditions of our model. In order to model a realistic population we use census data from the Irish Central Statistics Office (CSO, 2014a) to determine counts of people in each age group. Furthermore, vaccination uptake statistics for the whole of Ireland and by HSE region are available on Ireland's Health Protection Surveillance centre website going back to 1999 (HPSC, 2017). Another relevant data source was the Organisation for Economic Co-operation and Development (OECD) which reports vaccination rates for Ireland back to 1983 (OECD, 2017). This data is used to determine the proportion of people in each group who have been vaccinated and thus have an immunity to the disease.

3.2 Model

The equation based model is a compartmental model with age groups that determine vaccination status. There are 29 age groups in the model: one age group for each age between 0 and 27 and then another group for all other adults. When

applying the model to a particular town the size of the population within each compartment is created to match with the vaccination and population data. For each age group there are two equations: susceptible but not vaccinated (see Equation 3.1), and susceptible and vaccinated (see Equation 3.2). We consider homogeneous mixing within the model so the exposed, infected and recovered groups contain individuals from all 29 age groups. In total the model consists of 61 equations: two equations modelling the susceptible population for each of the 29 age groups (Equations 3.1 and 3.2), and three further equations that model the overall exposed, infected and recovered populations (Equations 3.3, 3.4, and 3.5):

$$\frac{dS_i}{dt} = \frac{-\beta S_i I}{N} \tag{3.1}$$

$$\frac{dS_{vi}}{dt} = -(1-\alpha)\frac{\beta S_{vi} I}{N} \tag{3.2}$$

$$\frac{dE}{dt} = \Sigma(\frac{\beta S_i I}{N} + (1-\alpha)\frac{\beta S_{vi} I}{N}) - \sigma E \tag{3.3}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{3.4}$$

$$\frac{dR}{dt} = \gamma I \tag{3.5}$$

Where S_i is the susceptible, unvaccinated population for age group i, S_{vi} is

the susceptible, vaccinated population for age group i, E is the exposed but not infectious population, I is the infectious population, R is the recovered population, N is the total population, β is the rate of transmission per unit time, α is the vaccination success rate, $\frac{1}{\sigma}$ is the duration of the exposed period and $\frac{1}{\gamma}$ is the duration of the infectious period. Using the measles dynamics discussed in Section 2.4, we take the duration of the exposed period to be 10 days and the duration of the infectious period to be 8 days. We then determine β using the formula $\beta = R_0\gamma$ where R_0 is the basic reproduction number, introduced in Section 2.1 (IDMDocs, 2019). This is a rearranged form of the equation presented earlier in Section 2.1.

The model is solved using Matlab ODE solver MATLAB (2017). The initial conditions for each town are set so that population in the 58 susceptible equations for age groups by vaccination status match the 2011 Irish census data and Irish vaccination statistics, there is one infected individual and there are no exposed or recovered individuals.

3.3 Model Evaluation

In the following sections we describe the evaluation and testing of the model. We will analyse the course of simulated outbreaks, by examining the infection curves created by the model, and then we run a sensitivity analysis looking at different levels of immunity in the population. The infection curve will help us to determine if the model is following the expected pattern for an SEIR model. The sensitivity

analysis will help to determine if the model acts as expected when parameters in the model are adjusted. For example, as vaccination rates, and thus the number of immune agents, increase the resulting size of the outbreak, and the chance that an outbreak will occur, should decrease. The model is further tested by comparing the results of our model to data from a real world outbreak that occurred in Schull, Ireland in 2012.

3.3.1 Modelling disease dynamics

When modelling an infectious disease, one important test is to determine if the model is capturing the correct disease dynamics. The dynamics of an SEIR model should roughly follow the curves in Figure 3.1. The SEIR infection curve plots the number of agents in each of the four categories – susceptible, exposed, infected and recovered at each time step.

As the curves in Figure 3.1 are produced using an SEIR differential equation model, we expect the results from our model to be similar. However, our model is broken down into a number of different compartments by age and vaccination status. Because of this it is necessary to make sure that the overall population dynamics of our model are as expected. The curves are done with a model with no vaccination or previous immunity so that everyone in the population is susceptible. So as to not have 58 different susceptible curves and to make the plot comparable to Figure 3.1, the output from the 58 susceptible curves are added together so all susceptible output is represented by one curve. The results can be seen in Figure

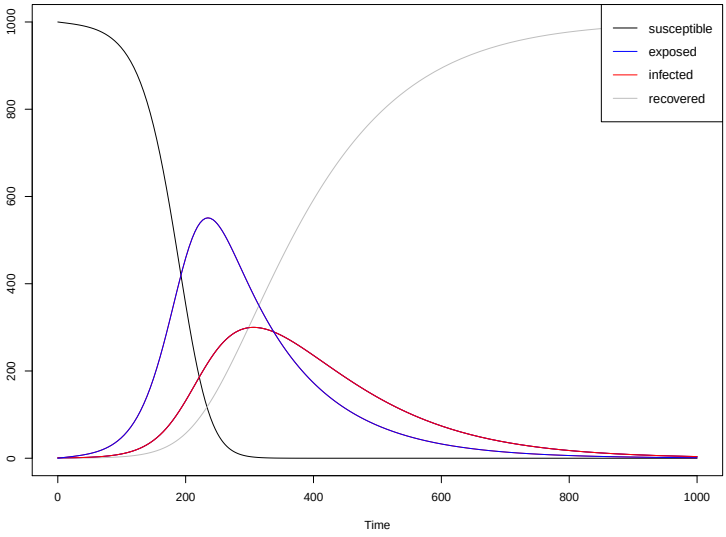


Figure 3.1: Example SEIR curves. The curves are generated using a basic SEIR differential equation model.

3.2.

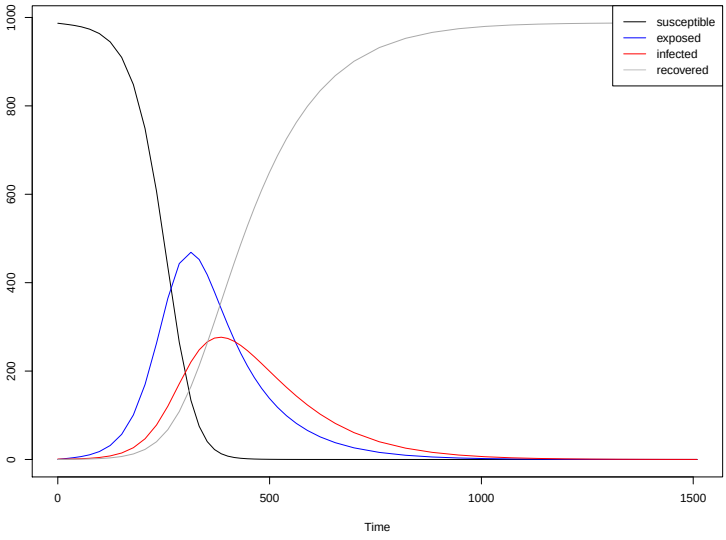


Figure 3.2: SEIR curves from the Equation Based model.

Comparing Figures 3.1 and 3.2, it can be seen that the two curves are nearly identical. Thus it can be assumed that our equation based model for the spread

of measles is accurately capturing the dynamics of the disease.

3.3.2 Sensitivity Analysis

After checking the model to make sure it produced the expected disease dynamics, a sensitivity analysis was run to determine if the model responds as expected to changes in different parameters, in particular we tested the models response to vaccination rates. For the analysis, the model is run on the town of Schull, Ireland. Schull is a small town with a population of about 1,000 and the town spans an area of about 17.1 km^2. Approximately 30% of the population in Schull are students and 11% of the population is not vaccinated or immune to measles. The initial conditions of the model are set to match the census data for Schull, Ireland. To analyse the results of the model we look at the change in the total number of people infected after the model has run (the model runs until there are no longer any individuals in the infected compartments). We look at a number of different scenarios with vaccinations and immunity: no individuals are vaccinated or immune, every individual is vaccinated or immune, or the percent of individuals vaccinated by age group is equal to the herd immunity rates. Herd immunity is discussed in Section 2.1 and refers to the concept that there is a critical number of individuals that need to be vaccinated or immunized to interrupt the transmission of an infectious disease in a population. Using the formulas in Section 2.1 for measles, we take R_0 (basic reproductive number) as 12 and V_e (vaccine effectiveness) as 97% (Nelson & Williams, 2007), which gives a critical vaccination

coverage level of 94.5%. The results for the three vaccination scenarios can be found in Table 3.1.

Immunity Level	Total Infected
No Immunity	987
All Immune	1
Herd Immunity	2

Table 3.1: Vaccination Scenarios Sensitivity Analysis for the Equation Based Model

The results show what was expected to occur giving further evidence that the model is working to capture measles dynamics accurately. When there is no immunity present in the model, the disease spreads through the entire population but when everyone is previously immune to the disease only the initial case is infected and it does not spread. Finally, in the herd immunity model, a second individual is infected but the outbreak does not spread beyond that individual. This again is what should occur in a population with herd immunity, one or two individuals may become sick but the infectious disease should not spread widely among the population.

3.3.3 Case Study: Schull 2012

While it is important to understand that the model behaves as expected it is even more important to show that the model can replicate an actual outbreak of measles. To test this, we present a case study that focuses on simulating a real measles outbreak that occurred in Schull, Ireland in 2012 (HSE, September 2012).

Over a three-month period starting on 9 April 2012 and ending 15 June 2012

there were 63 cases of measles notified in West Cork (the region of Ireland in which Schull is situated in) (HSE, September 2012). The initial focus of the outbreak was on the Schull and Skibbereen areas but as the outbreak went on cases were notified throughout West Cork including Bantry, Bandon, Dunmanway and Clonakilty (HSE, June 2012). Seventy-eight percent of the cases were in the 10-19 year age group, and only four cases were in the under-five age group. Lack of vaccinations is thought to have played a large part in the outbreak. Vaccination status is known for 59 of the 63 cases: out of those 59 cases 55 were unvaccinated (HSE, September 2012). According to the Health Services Executive in Ireland, at the time of the outbreak the uptake of the MMR vaccine at age two years was 92% nationally, however, in West Cork the uptake was 86%: West Cork has historically been an area of low uptake of vaccinations. In response to the outbreak, it was recommended that infants aged 6-12 months resident in the area have an early dose of the MMR vaccine, and unvaccinated siblings of infected individuals were asked to self-isolate for the duration of the incubation period. Of the 63 cases in the West Cork region approximately 30 cases were in Schull (HSE, June 2012).

Similar to the sensitivity analysis in the previous section: the initial conditions are set so that each age category in the model matches the census data from Schull Ireland. To determine the proportion of each age group that is in the vaccinated category we use West Cork vaccination rates for MMR.

When the equation based model is run we find that the total number of individuals who are in the recovered category in the model, those individuals who

were infected during the model and have since recovered is 29.868. Which is the number of cases that were in Schull, Ireland during the outbreak. This can be taken as evidence that our model is predicting close to what occurred in the actual outbreak.

3.4 Simulating Additional Towns

In order to show that the model can not only model a single town but can also be applied to any other town, we simulate a measles outbreak in 33 different towns. The towns are small to medium size towns in Ireland with populations between 390 and 9,548. The towns included were selected to have a range of diverse towns to test the model.

The results for each town along with town characteristics such as population size can be seen in Table 3.2. For each town the equation based model was run changing the initial conditions to match the correct number of people in each age category for the town based off of the CSO census data. To reduce variability from the different levels of vaccinations based on regions in Ireland the vaccination rates used were the all Ireland rates. As the vaccination rates are by age, there may be variability in the total vaccination rates of a town if the age structure differs. For example, an older population should have higher levels of vaccination and immunity than a younger population.

From the results we can see that our model is able to simulate a measles outbreak in a number of different towns and produce different results based off of

the different population structures. In addition, we can see that for towns that are similar in population size in some cases the equation based model produces similar results such as for Schull and Strokestown. Both towns have populations of approximately 1,000 and the equation based model results have 12.57 infected at the end of the model for Schull and 13.14 in Strokestown. However, there are some cases where similar towns have different equation based results, for example Shanagolden has a similar population to both Schull and Strokestown but has 7.26 infected at the end of the model. Another such example can be seen comparing the towns of Ardamine and Rathnew. The two towns have nearly identical populations, 3,293 and 3,294 respectively but the model results in 44.80 total infected for Ardamine and 85.89 infected for Rathnew. The differences in results is likely due to the model capturing some interactions between the population size, the age structure and the vaccination rates of the towns. This is a desired result as typically multiple factors influence the course of an infectious disease outbreak.

Town	Population	Population Density	Total Infected	Percent of Population
Arainn	1,251	26.35	10.65	0.85
Ardamine	3,293	141.15	44.80	1.36
Ardfert	997	125.09	24.60	2.47
Arranmore	514	28.43	5.40	1.05
Bagenalstown	3,421	190.06	52.80	1.63
Ballyjamesduff	3,134	145.09	106.33	0.34
Banagher	1,993	100.40	38.05	1.91
Blarney	5,310	227.90	82.27	1.54
Castlereagh	3,077	76.75	14.90	0.48
Clane	7,527	398.46	219.75	2.92
Croom	1,690	93.01	62.33	3.69
Donegal	4,010	127.34	35.97	0.90
Gort	2,671	238.27	38.17	1.43
Kenmare	2,912	52.36	14.81	0.51
Kilcock	6,234	385.61	162.07	2.60
Kildare	9,325	251.42	259.71	2.79
Kilkee	1,037	197.15	8.16	0.79
Killadysert	922	14.42	9.61	1.04
Kinsale	6,871	530.17	22.92	0.33
Lisdoonvarna	861	66.44	12.08	1.40
Louisburgh	983	42.19	11.50	1.17
Moate	3,046	149.75	55.86	1.82
Oranmore	4,325	193.25	90.72	2.10
Portmagee	390	23.26	6.17	1.58
Rathnew	3,294	477.39	85.89	2.61
Roscrea	6,318	130.40	146.30	2.32
Rosslare	2,057	114.91	11.01	0.54
Roundstone	459	16.39	15.57	3.39
Schull	987	57.96	12.57	1.27
Shanagolden	946	53.18	7.26	0.77
Stamullin	4,694	124.58	95.05	2.02
Strokestown	1,003	55.38	13.14	1.31
Tramore	9,548	575.18	141.75	1.48

Table 3.2: Population and model results for each of the 33 selected towns

3.5 Conclusion

The analysis of our equation based model for measles spread shows that an equation based model is able to accurately capture expected disease dynamics that are sensitive to factors such as vaccination rates. We also show that such a model can be created to model a real world outbreak with our model reaching the same number of infected individuals as in the outbreak of measles in Schull, Ireland in 2012 and that the model is able to model a set of 33 different towns by changing the initial conditions and the results vary based on the towns and those initial conditions.

However, the model as it stands produces a single outbreak scenario for each town which does not take into account day to day variability. While we may have produced similar results to a real world outbreak, the course of that outbreak occurred as it did because of interactions between not just the characteristics of the town but also the actions of the individuals in the town. One individual making different decisions could have drastically changed the outbreak. In addition, adding any more factors into the model will require adding additional interaction terms which will make the model more complicated and more difficult to solve.

Equation based models can be considered a "top" down method of modelling at the population level instead of the individual level. Each compartment or equation in the model is homogeneous and any new characteristic in the model requires additional equations. "Top down" methods allow for quick model run time and easy scalability, however, they do not allow for important individual factors

that might shape an outbreak. Thus, while our model is accurate in simulating a single outbreak, it is important to investigate other methods of modelling infectious disease spread.

Chapter 4

Taxonomy of Agent-based Models for Infectious Disease Epidemiology

When we began the research programme reported in this thesis there was no standard for creating an agent-based model to study the spread of an infectious disease through a population. Therefore, we developed a taxonomy of agent-based models for epidemiology based on a review of the literature. In order to understand a simulation and its potential uses it is important to note how the components are combined. A taxonomy of epidemiological simulation models based on the level of specificity of the disease, society, transportation and environment model can be created to aid in the classification of agent-based models for human infectious diseases. A taxonomy can help researchers understand where to place their model

in the body of existing work and can help them choose what level of complexity they need in their model. We aim to use the taxonomy to guide the creation of our own model. The taxonomy was first presented in Hunter et al. (2017).

Figure 4.1 illustrates the taxonomy we propose. The models are first broken down into two categories: general and specific disease models. Those two categories are then broken down to general or specific society simulations, if the model includes a transportation model, and the environmental factors in the model. The grey branches and boxes of the taxonomy that are outlined in grey are those combinations of component types that we did not find in our literature review and based on an analysis do not think would be feasible combinations. For example, we do not feel it would makes sense to have an environment made up of maps or maps plus other factors if the model has a general society model. This is because the maps that would be used would be based on the society that is being modelled and in this case no specific society would be modelled. The following sections will describe the different components in the taxonomy. Note that whereas Sections 2.3.1 through 2.3.4 introduced and reviewed the design spaces of the different components of an agent based model for infectious diseases, Sections 4.1 through 4.4 describe our taxonomy based on those components.

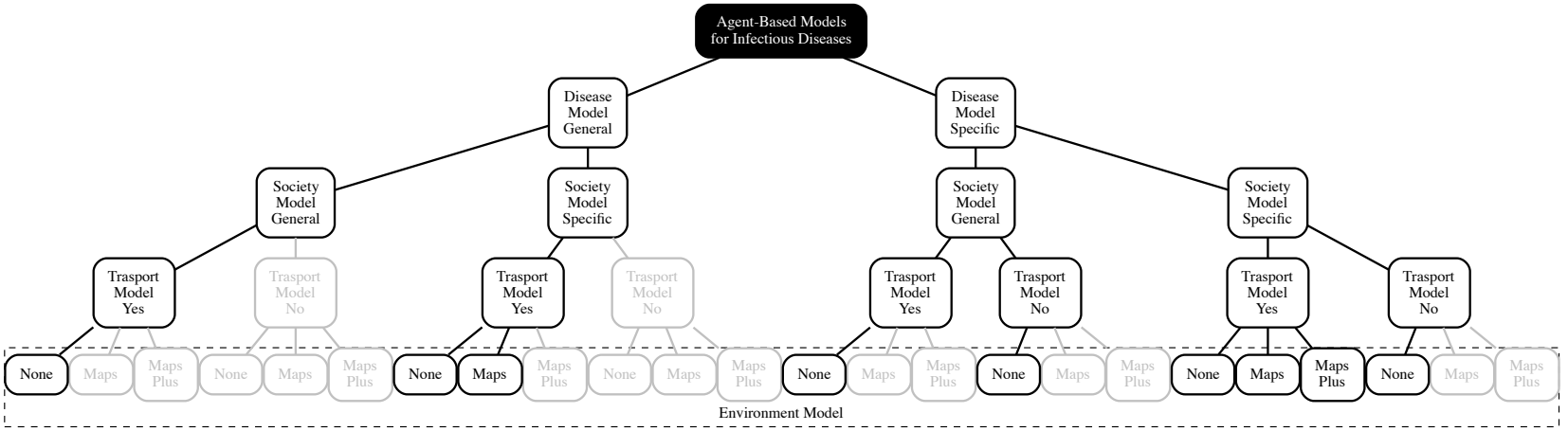


Figure 4.1: Taxonomy of Epidemiological Agent-Based Models.
Grey branches and boxes outlined in grey are those combinations of component types that we did not find in our literature review and based on an analysis do not think would be feasible combinations.

4.1 Disease Model

The taxonomy as seen in Figure 4.1, breaks the disease model component of a simulation down into *specific* versus *general* disease models. General disease models are those that use basic SIR disease dynamics in the model with parameters that can be adapted for various airborne diseases. Simulations with a general disease model prove useful for planning and creation of policy as only one simulation needs to be created to analyse the effects of different outbreaks. The general disease model also allows researchers to save time if they want to switch their model from showing the effects of an influenza outbreak to a measles outbreak or any other disease with similar transmission dynamics. This is the case with Barrett et al. (2008)'s Episimdemics, Aleman et al. (2011)'s model for Ontario and Lee et al. (2008) model for Alleghany County, Pennsylvania.

Often a model will focus on a specific disease because there is some reason that a general model will not capture the disease spread accurately enough. This can be the case if the infection dynamics of a disease are not typical, for example Ebola can be transmitted at funerals and cholera is transmitted through drinking water. Agent-based models have been based on specific outbreaks such as the Ebola outbreak in Liberia. Not only does this model include specifics to how Ebola spreads, such as contact at funerals, but the model is specific to Liberia including the number of hospital beds that were used for Ebola patients over the course of the epidemic (Merler et al., 2015) However, many of the agent-based models that focus on a specific disease model influenza. The transmission dynamics of

influenza are closer to a general model than some other diseases such as TB or malaria and as such a SIR type model can be used. These models either focus on a specific strain of influenza such as H1N1 or H5N1, or treat influenza generally (Friás-Martínez et al., 2011; Dibble et al., 2007; Rakowski et al., 2010a). Specific agent-based models can also be used to determine how given interventions affect the spread of a virus. Among other topics agent-based models have been created to determine the effects that the government mandates had on the spread of the H1N1 virus in Mexico and how vaccination programs affect the incidence rate of Human papillomavirus (HPV) in Denmark (Friás-Martínez et al., 2011; Olsen & Jepsen, 2010).

4.2 Society Model

The next stage in the taxonomy is the society model. Society models can be described as *specific* if they were created to model an actual society using real data and *general* otherwise. It is most often the case that a specific disease model is paired with a specific society model, as the idea behind such a model is typically to capture the dynamics of a past or current outbreak. In order to do this it is not just necessary to accurately model the disease but the society as well. However, there are cases where a specific disease is paired with a general society model. This would occur in cases where the model might be used for disease dynamics research where an overly realistic society is not needed such as in Duan et al. (2013)'s model looking into the possibility of identifying super spreaders.

In some cases, such as with Barrett et al. (2008)'s EpiSimdemics, a general disease model is integrated with a specific society model. Such a model would be used as a public health tool where the effects of any possible new outbreak can be modelled on the given society. These models are most often used for planning for future outbreaks or epidemics as a public health tool for decision making. EpiSimdemics is a model that was created to scale to social networks with 100 million individuals. The parameters of the model can be altered in order to model different infectious diseases. Another modelling tool was created for the Greater Toronto area in Canada to determine the best mitigation strategies in the case of a potential epidemic (Aleman et al., 2011).

General disease models can also be combined with a general society model. Although the results of the models cannot be directly applied to a given society, these types of models can be used for research purposes. For example, models can show how factors such as fleeing from fear might change the course of the outbreak and how the differences in the number of individuals in different categories (susceptible, infected and recovered) over time compared to the results of an equation-based model will broadly show the effect that spatial mixing has on the outbreak of a disease. The models by Epstein et al. (2008) and Dunham (2005) are both examples of these types of models.

4.3 Transportation Model

We consider two levels of transportation for the taxonomy: models with transport and models without transport. Although there could potentially be finer levels created based on the complexity in the transportation model we felt that the boundaries between these levels were too fuzzy to be useful.

Models without transport tend to be matched with specific disease models as not including transportation in the model is most useful when the disease dynamics can be modelled with contact network structures versus day to day interactions. This is often the case with blood/bodily fluid borne diseases such as HIV or HPV, for example Olsen & Jepsen (2010)'s model for the spread of HPV in Denmark. In such cases adding transportation into the model would only serve to slow the simulation down. Models without transport can be paired with either a specific or general society based on the aim of the model.

Models with transportation are paired with both general and specific disease models. The types of diseases that are transmitted with airborne transmission that can be substituted into a general model are often relatively contagious and a transportation model helps to better capture the spread of the disease by identifying random contacts outside of an agent's family or friends. Epstein et al. (2008) use a general disease model and a transportation model to determine how agents fleeing during an epidemic will lead to greater spread of the disease. Similarly transportation can help capture the dynamics of disease spread for models of specific diseases. For example, Crooks & Hailegiorgis (2014) use agents' movements to

determine if, when and where an agent is drinking contaminated water that may lead to a cholera infection. Vector borne diseases can also be reliant on movement, if an agent travels to an area with a higher concentration of the vector population it will be more likely that the agent will be infected (Linard et al., 2009). Pairing a transport model with a specific disease model helps to capture dynamics of many infectious disease that may have been modelled incompletely without movement.

Similar to the models with no transport, models with transport can be matched with either specific or general society models. The choice of society would be determined by the researchers and their goals for the model. For example, Olsen & Jepsen (2010) wanted to study the effects of the HPV vaccine on the population of Denmark. To do this they needed to include a specific society because they wanted their results to be specific to Denmark, however, because HPV is a sexually transmitted disease the model does not need to include transportation as daily interactions on the road or in work or school will not spread the disease.

4.4 Environmental Model

The taxonomy breaks the environmental model down to three levels: no added environment, maps, and maps plus other environmental factors. Other environmental factors could be anything from temperature and precipitation to a vector population or a hydrology model. Models with other environmental factors added into the model are usually combined with a specific disease model since the environmental factors added into the model should be factors that are related to the

disease. For example, the model by Crooks & Hailegiorgis (2014) is a model for the spread of cholera in a specific refugee camp and includes hydrology models and precipitation models as they are essential to the spread of the disease. In most cases these models are also paired with a specific society since the environmental factors are based on what is seen in the real world. Models with environmental factors are also typically paired with a model with transportation. A simulation with specific disease, specific society and high level environment would create a simulation where the results can be easily applied to a real life scenario. However, the models will also be data heavy which could lead to slow initialization and computing time.

Environmental models that include roads, buildings and/or maps are nearly always paired with a specific society model. This makes sense in the idea that in order to add roads or a map in the environment the researchers would need to choose which map or which roads to use based on the society that is being modelled. Additionally, models with an environment that is made up of maps or roads will usually include transportation. There is not much point in creating an environment with roads if the agents do not move along them. Such models are matched with either a general or a specific disease model.

Models with no added environment are those where the simulation is solely made up of agents interacting with each other in an open space. These types of environments can be in simulations that do not include transportation but focus on a specific disease, such as Olsen & Jepsen (2010)'s model. Because the agents

will not move through the environment there isn't as much of a need to create any detail in the environment for the agents to interact with. Other models with no added environment will include transportation. Most often these models are dealing with a general society and either a specific or general disease model.

4.5 Applying the Taxonomy

Table 4.1 puts the 20 models analyzed in our literature review into the classification system. If the disease model or the society model is a specific model the name of the disease or society is also included in the table.

The table also includes the possible use of the model. Based on the models reviewed, there are four main uses for agent-based models of infectious diseases: disease dynamics research, agent-based modelling research, epidemic planning, and lessons learned. Disease dynamics research focuses on learning information about how a disease will transmit in a circumstance that would otherwise be hard to learn without a real life outbreak scenario. For example, how finding and treating super-spreaders can help to lessen an epidemic and how effective contact tracing can help stop outbreaks of TB (Duan et al., 2013; Tian et al., 2013). Agent-based modelling research is concerned with finding new ways to use agent-based models and new methods for creating agent-based models. For example, Bobashev et al. (2007) explore how to combine agent-based and equation based models. A model used for epidemic planning such as Barrett et al. (2008)'s Episimdemics model is created to learn the best strategies to deal with outbreaks prior to an outbreak

occurring while lessons learned is the idea of modelling a past outbreak in order to learn from what happened in the past and to determine if the measures taken to stop the spread of the disease were successful. For example, Friás-Martínez et al. (2011) modelled the 2009 H1N1 outbreak in Mexico City in order to determine if restrictions on movement affected the course of the epidemic.

Using the table to find similarities in the models that have the same use can help to better use the taxonomy. For example, determining that a lessons learned model always contains a specific disease and specific society will direct researchers to that branch on the taxonomy and help them to make decisions on the other components they need to include in the model.

Models that focus on a specific society with either a general or specific disease tend to be used for epidemic planning. A number of the models reviewed, (Crooks & Hailegiorgis, 2014; Rakowski et al., 2010a), attempt to accurately simulate the spread of a specific disease so that the model could be used in the future to determine best practices. Other models such as, Barrett et al. (2008) and Aleman et al. (2011) use a general disease to create a model that can be used for multiple outbreaks. Results published from a study using the Simdemics model show that a combination of school closures, individual adaptive behaviour, and targeted antiviral distribution could reduce the impact of an influenza-like pandemic by 87% and the income loss from such a pandemic would decrease by 82% compared to a base case (Apolloni et al., 2009).

Paper	Disease	Society	Transport	Environment	Use
Ajelli et al. 2010	General	Specific (Italy)	Yes	Maps	Agent-Based Model Research
Aleman et al. 2011	General	Specific (Ontario)	Yes	Maps	Epidemic Planning
Barrett et al. 2008	General	Specific (USA)	Yes	Maps	Epidemic Planning
Bobashev et al. 2007	General	Specific (World)	Yes	No	Agent-Based Model Research
Crooks and Hailegiorgis 2014	Specific (Cholera)	Specific (Dadaab refugee camp)	Yes	Maps plus	Epidemic Planning
Dibble et al. 2007	Specific (H5N1)	Specific (USA)	Yes	None	Epidemic Planning
Duan et al. 2013	Specific (SARS)	General	Yes	None	Disease Dynamics Research
Dunham 2005	General	General	Yes	None	Disease Dynamics Research
Epstein et al. 2008	General	General	Yes	None	Disease Dynamics Research
Frias-Martinez et al. 2011	Specific (H1N1)	Specific (Mexico City)	Yes	Maps	Lessons Learned
Lee et al. 2010	General	Specific (Allegheny, PA, USA)	Yes	Maps	Epidemic Planning
Linard et al. 2009	Specific (Malaria)	Specific (South France)	Yes	Maps plus	Epidemic Planning
Mao 2014	Specific (Influenza)	Specific (Buffalo, NY, USA)	Yes	Maps	Disease Dynamics Research
Merler et al. 2015	Specific (Ebola)	Specific (Liberia)	Yes	Maps	Lessons Learned
Olsen and Jepsen 2010	Specific (HPV)	Specific (Denmark)	No	None	Epidemic Planning
Perez and Dragicevic 2009	General	Specific (Burnby, Canada)	Yes	Maps	Epidemic Planning
Rakowski et al. 2010	Specific (Influenza)	Specific (Poland)	Yes	Maps	Epidemic Planning
Simoes 2006	Specific (Mumps)	Specific (Portugal)	Yes	Maps	Lessons Learned
Skvortsov et al. 2007	General	Specific (Australia town)	Yes	Maps	Epidemic Planning
Tian et al. 2013	Specific (Tuberculosis)	Specific (Saskathewan, Canada)	No	None	Disease Dynamics Research

Table 4.1: Simulation Classification Table.
The *disease*, *society*, *transport*, and *environment* columns place the papers in our taxonomy while the *use* column details the intended uses of the model. The use of the model although not part of our classification system, can be affected by where a model falls in the taxonomy.

The EpiSimdemics model is able to simulate detailed information on a disease spreading through a population including the individuals infected, where they were infected and who infected them. The information EpiSimdemics provides allows for identification of the severity of the epidemic as a whole and in certain subpopulations. The model has been used for multiple studies including those on pandemic planning for the US Department of Defence and the US Department of Health and Human Services. Looking at the effects of sequestering military subpopulations during a pandemic, the EpiSimdemics model determined that counter-intuitively sequestration may lead to more infections. It was determined this was because certain diseases can be infectious before being symptomatic and although overall contacts would decrease with sequestration contacts in a smaller group of individuals, those who were sharing military quarters, would increase: resulting in infectious individuals being in close contact with susceptible individuals for a long period of time (Barrett et al., 2008). The general disease model combined with specific society and transportation models allows for the user of the model to change the infection dynamics based on what situation they would like to study. This saves the effort of recreating a model for the same population every time a study needs to be done and gives the user the advantage of having a previously validated model. Another similar planning result obtained from a specific disease model is a cost-effectiveness analysis. Olsen & Jepsen (2010) use an agent-based model to determine cost-effectiveness ratios for HPV vaccinations and determine that while a new vaccination program will incur costs, in the long term it will save

treatment costs and improve quality of life and survival.

If a researcher wished to create a model for epidemic planning they could go to the taxonomy and look at the branches that contain specific society. Looking at the taxonomy, if they also wished to include a general disease model they would know that a transportation model should be included and they would only need to decide on no added environment or maps. Alternatively if they wanted to include a specific disease model, the researcher might need to decide if they wanted to include transportation in their model based off of the transmission of the disease being modelled (human-to-human, food or water to human, vector to human). Deciding to not include transportation would also result in not including any added environment, while deciding to include transportation would require a decision on what level of environment would need to be added.

From the table, models that are for lessons learned from a past outbreak tend to be created with a specific disease and specific society model. For example, Frias-Martinez created a model for the H1N1 outbreak in Mexico city in order to evaluate the cities mitigation strategy (Friás-Martínez et al., 2011). Similarly Merler et al. (2015), looked into the Ebola outbreak in Liberia to determine if safe funerals and Ebola patient only medical centers affected the outbreak. Knowing this, if a researcher wanted to create a model that would be used for lessons learned they could go to the taxonomy and follow the branch to specific disease and specific society. Based on the disease being modelled and the type of transmission, they could then decide if the model should include transportation and how much added

environment to include.

Models that have a result focused on agent-based model research are often created to find a solution to some of the problems in the field of agent-based modelling for infectious disease epidemiology. One of the main barriers in the uptake of agent-based models is the time it can take to run a detailed simulation coupled with the large amount of processing power needed. In order to overcome this barrier experimentation must be done to create more efficient agent-based models. This is already occurring in cases such as Bobashev et al. (2007) where an agent-based model is combined with an equation based model to improve efficiency. As agent-based models become faster and more efficient, more detail will be able to be added to the models. Hopefully this will result in larger uptake of agent-based models to help determine policy and direct research. These types of models will usually have a general disease, which would then require the modellers to decide if they wanted a general or specific society and if they choose a specific society if they should add environment to the model.

Models that are created to look into disease dynamics research can be placed anywhere on the taxonomy. Results from disease dynamics research models include learning about the effects of super-spreaders on an outbreak (Duan et al., 2013), looking into the effects of fleeing an outbreak (Epstein et al., 2008), or what effect the spread of fear and knowledge of the outbreak will have Mao (2014). As these models can have a general or specific disease, a general or specific society, transportation or no transportation and any level of environment, to decide what

will work best for their model a researcher should go through each layer of the taxonomy and determine what will work best for them. For example, to see the effects of fleeing on an outbreak, Epstein et al. (2008) decided on a general disease model as they were not focusing on a specific disease but wanted to look more generally at what happens and a general society again to see the general effects of fleeing that might occur in any society. Once a general disease and society were chosen the options on the taxonomy include transportation and no added environment.

Although we have fit all of the models we reviewed into our taxonomy we are aware that no taxonomy can be completely comprehensive and there may be models that do not fit nicely into our classification. Even if this occurs we feel that our taxonomy is still a useful tool as it will work for the majority of agent-based infectious disease epidemiology models and was created based on evidence from the literature. It should also be noted that the taxonomy is not all inclusive and that there are other characteristics of the simulations that could be included. The taxonomy should, however, aid readers and simulation designers alike in determining the use of a simulation based on the different components of the simulation are handled and what to expect from the results.

4.6 Conclusions

Agent-based models can be a useful tool in helping to stop or prevent the spread of an infectious disease. Models such as Barrett et al. (2008)'s Episimdemics are already being used to influence policy. Merler et al. (2015) and Friás-Martínez et

al. (2011) have studied past outbreaks to determine the success of interventions to help inform in case of future outbreaks. However, the freedom and flexibility in agent-based model design allows many different type of models to be created even just in the field of infectious disease epidemiology. Yet the lack of clear protocols in creating and describing agent-based models can lead to confusion in understanding the methodology of a given agent-based model. Because of this it is essential to understand the different types of agent-based epidemiology models and how they relate to each other. The literature shows that similarities among existing agent-based infectious disease epidemiology models exist and that there are different ways to compare the simulations. These comparisons tend to be driven by similarities or differences in the components of the model, disease, society transportation and environment, and how the model handles the components.

For both disease and society we found that models in the literature tend to create either specific components based on data or general components where parameters can be adjusted to model multiple diseases or results can be applied to any society. The choice of general or specific disease model or a general or specific society model will have an effect on the transportation and environment components used, advantages and disadvantages of the model, possible uses of the model and the validation process.

As there are many possible combinations of the disease, society, transportation, and environment components of a model, each with potentially different uses, validation techniques, advantages and disadvantages we felt that the current liter-

ature was missing a classification tool. Using the knowledge we gained from our literature review we formulated our taxonomy. The taxonomy should aid readers and modellers alike in determining the use of a model based on how the different components of the simulation fit together. One of the problems with the current agent-based modelling literature is the lack of clear definitions and standards for agent-based models in infectious disease epidemiology and the components of those models due to the flexibility and freedom allowed in model design. We feel that creating a taxonomy can help to classify agent-based infectious disease epidemiological models and is a move towards solving the problem of definition without sacrificing the flexibility that attracts researchers to the agent-based modelling field.

In addition to helping classify the existing models in the literature we feel that the taxonomy can help researchers in creating models through determining which components are necessary for their intended use. For example, if a model is being created for epidemic planning it will need to have a specific society component. Once the components of a model are determined the taxonomy can help researchers identify the available methods for validation. For instance, if a general disease model or general society is used it may not be possible to compare the results to past outbreaks, while the use of a specific disease and a specific society makes the use of past outbreaks as a validation method a possibility. We see this as a real benefit of the taxonomy. By helping researchers identify the range of validation techniques that are suitable for a specific model the taxonomy can help

standardise the approaches to validation that are used for agent based models for epidemiology. As discussed in Section 2.3.5, the fact that different researchers employ very different levels of validation of their models is a recognised issue in agent-based modelling research so something to help standardisation could be a real benefit. We used the taxonomy to guide the creation of our own agent-based model that will be presented in the next chapter. Aiming to create a measles model for the Irish context we want a specific society and a specific disease and can then use the taxonomy to help determine what level of transportation and environment should be included in the model.

Chapter 5

An Agent-Based Model of Measles in an Irish Town

We have already shown an equation based model for the spread of infectious diseases through Irish towns in Chapter 3 that is able to show differences between towns based off of population differences and differences in vaccination rate. In this chapter we create an agent-based model for the same Irish towns in order to investigate if an agent-based model is better able to capture the interactions between town factors and characteristics that will influence the spread of the infectious disease. We use the taxonomy outlined in the previous chapter to help guide the creation of our model. Our aim is to create an agent-based model with specific disease and society components that includes transportation and maps. Section 5.1 describes the data used to create the model. Then in Section 5.2 we describe the model, breaking it down into parts based on the taxonomy from Chapter 4,

evaluate the model using sensitivity analysis and a case study in Section 5.3 and finally show results simulating 33 different towns in Ireland in Section 5.4. The model and results were previously published in Hunter et al. (2018b).

5.1 Data

Creating the agent-based models described in this chapter required a variety of data to be aggregated, prepared and analysed including population statistics, GIS data, school and workplace locations and vaccination data. The majority of data used comes from Ireland's Central Statistics Office (CSO)(CSO, 2014a), but other sources are also used. We use only publicly available open data sources to create the model which leads to greater reproducibility. The growth of big data and more data sets becoming openly available allows for the creation of more detailed agent-based models. Governments are making data sets more widely available, allowing anyone to have access to data sets on topics such as population, health, economics and transportation. Often the links to such data sets are being made easily accessible on one platform. For example, Ireland's open data portal (data.gov.ie) or the city of Glasgow's open data website (data.glasgow.gov.uk). The more realistic a model is to the society in question the easier it is to apply those results to real world scenarios. Openly available data has the additional advantage of reproducibility as anyone has access to the data to recreate the model or update the model with new data. In addition, although our model is tested on the Irish context it is easily portable between towns in Ireland and if the same level of data exists for a town

in a different country our model could be used to simulate an outbreak for that town. The following sections outline the sources of the data used in the model.

5.1.1 Population statistics

Population statistics are used within the model to create a specific society of agents. Real data is used to determine the age and gender breakdowns of our populations along with correct distribution of household size and other household characteristics such as child age. The CSO provides a wealth of open access data. The data is taken from the results of the Irish census which occurs every five years. The data used for our model is from the 2011 Irish census, data from the 2016 census has recently been made available, however the 2011 data is more suitable for the outbreak we attempt to simulate as it occurred in 2012. The census data is organized into fifteen different themes each with a set of tables containing information on the population of Ireland under that theme. The themes are described in Table 5.1. The themes used to create the model are theme 1, theme 4, theme 5 and theme 8.

Theme 1: Sex, Age and Marital Status	Theme 9: Social Class and Socio-Economic Group
Theme 2: Migration, Ethnicity and Religion	Theme 10: Education
Theme 3: Irish language	Theme 11: Commuting
Theme 4: Families	Theme 12: Disability, Carers and General Health
Theme 5: Private Households	Theme 13: Occupation
Theme 6: Housing	Theme 14: Industries
Theme 7: Communal Establishments	Theme 15: PC and Internet Access
Theme 8: Principal Status	

Table 5.1: The 15 themes from the CSO census data tables CSO (2014b)

Data can be downloaded at multiple geographic levels, the lowest being the small area (CSO, 2014a). Small areas are areas of population that contain between

50 and 200 dwellings. We base our simulations on data at the small area level. The CSO make available a data set (delivered in csv format) for all small areas in Ireland containing data for each table within each theme. When simulating a specific town the small areas related to that town and the necessary tables can be selected from the data set. The small area boundary file discussed in the next section provides a mapping between small areas and towns.

5.1.2 GIS data

Various sources of GIS data are used in our models. GIS data not only gives us the town boundaries but also residential, commercial and recreational areas within the town that help to define where the agents live, work, and travel.

The CSO provides access to boundary files from the 2011 census. The files contain the boundaries at different levels including provinces, counties, electoral divisions, towns and small areas (CSO, 2014a). The data set downloaded from the CSO website contained small area information for all of Ireland: the QGIS (QGIS, 2009) software was used to select only the small areas that overlapped with the town being simulated so the data could be loaded into Netlogo. The small area boundaries do not always match town boundaries, thus the small area data set could potentially cover more area than the town being simulated.

Zoning data is taken from two sources: Open Street Maps (OpenStreetMap contributors, 2017) and Myplan.ie (Myplan.ie, 2017). Myplan.ie gives the shape files that include local area development plans. Open Street Maps provides land

use data. The land use data is a shape file that provides information on if the land is used for residential, commercial, retail or industrial purposes. The data set can also provide more detailed information such as if the land is used for religious purposes, sports pitches, cemeteries or reservoirs. Neither source is comprehensive and there are some areas in the towns for which zoning data is not available. The different zoning and land use types are sorted into six categories: open, town center, community, residential, commercial and mixed.

5.1.3 School locations

In order to determine both the number of schools in a town and their locations we use data from the Department of Education and Skills in Ireland. They provide data on individual schools, including enrolment and type of school (primary or secondary)(DOE, 2017). The data set includes the longitude and latitude of the schools. These are then geocoded in QGIS (QGIS, 2009) in order to create a GIS shape file that can be combined with the town boundary and land use shape files and loaded into Netlogo.

5.1.4 Vaccination data

Vaccination statistics are used to determine the number of agents in our model who have been vaccinated and thus are immune to the disease. Vaccination statistics for Ireland can be found on the childhood vaccination schedule. Statistics are presented for Ireland as a whole and broken up into Health Service Executive

(HSE) regions. Vaccination uptake statistics for the whole of Ireland and by HSE region are available on Ireland's Health Protection Surveillance centre website going back to 1999 (HPSC, 2017). The Organisation for Economic Co-operation and Development (OECD) reports vaccination rates for Ireland back to 1983 (OECD, 2017). When initiating the model the choice of all Ireland vaccination rates or vaccination rates for a specific HSE region based on the town being modelled must be made. Further discussion of how vaccination rates are used in the model can be found in Section 5.2.2.

5.2 Model

We use the computer software Netlogo (Wilensky, 1999a) to implement the simulations described in this thesis. However, the open-data driven approach described in this thesis could be used with any agent-based modelling tool or platform. Netlogo is an easy to use and popular environment for creating ABMs (Gilbert, 2008). It does, however, have disadvantages, one of the biggest being the speed of the program. When modelling simulations with a small number of agents Netlogo works well, however, once the number of agents gets large the simulation can become prohibitively slow.

The model used in the experiments described in this chapter can stimulate outbreaks of measles-like diseases in small to medium sized Irish towns. As it is a small, relatively simple model, the aim is validate this basic model and then in later chapters we will expand on this model adding detail and scale knowing

that the basic model dynamics are correct. Each agent in the model simulates the behaviour of a resident of the town. The simulations of the towns and the behaviours of the agents in the towns are driven by the data sets described in the previous section. The model we use in the experiments is outlined using the ODD protocol (Grimm et al., 2010) in Appendix A and can be found online in the Netlogo User Community Model Library (`http://ccl.northwestern.edu/netlogo/models/community/town_model_burnin`) The model can be adapted to simulate any town in Ireland or elsewhere as long as the data sets listed in Section 5.1 are available for the new towns. The following sections give a brief overview of the environment, society, transportation and disease components of the model, as well as the schedule used to simulate agent behaviour.

5.2.1 Environment

The towns are created in Netlogo. The Netlogo world is a two dimensional grid where the squares that make up the grid are referred to as patches. For the purpose of our model each patch has an area approximately equal to 111 m^2. If two agents occupy the same patch they are considered to be in the same place and in contact with each other. Small area data sets from Ireland's CSO are used to create the basic geographic layout of the town. Primary and secondary schools are added into the model using data from Ireland's Department of Education and Skills. Zoning data, as described in the previous section, is used to place households and work places in appropriate locations within the small areas and to

determine community locations. The population data for each small area specifies the number of households in that small area, and a corresponding number of households is added to each small area in the simulation. The population data specifies the distribution of household types in each small area *(single, couple, couple plus others, couple with children, couple with children plus others, single parent, singles parent plus others or other)* and the houses in our simulated small areas are assigned types randomly selected following this distribution. Household types are assigned concurrently with creating agents so that accurate household structures are created. Figure 5.1 provides an example of the model environment created for the town of Schull, Ireland. More detail on the creation of households can be found in the "Initialization" section of Appendix 1. The model is given a start week, which is used to determine if summer holidays occur over the lifetime of the model, and a year which is used to determine vaccination rates.

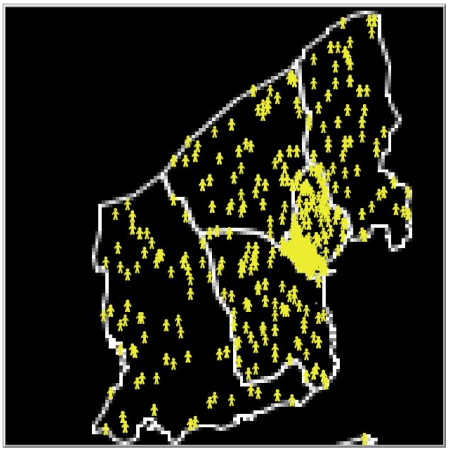


Figure 5.1: Example of Model Environment. The environment created by the model for the town of Schull, Ireland. The white lines are the boundaries of the small areas and the yellow agents are located at the agent households in the model.

5.2.2 Society

Agents are added to the town based on the population data described in Section 5.1.1, adults are added first and given an age and sex based on the distributions of age and sex in the appropriate small area from the census data. Children are then added into households that have a type that includes children and are given an age and sex based on the appropriate distributions from the census data. All agents are given an economic status *(student, retired, looking for first job, unemployed, sick/ disabled, stay at home, work)* based on the distribution in the relevant small area in the census data. Agents with the economic status of work are assigned to a random workplace within the town and agents with the economic status of student are assigned randomly to one of the schools of the appropriate level (*preschool*, *primary*, *secondary*) in the town. Irish vaccination data is used to determine the percentage of each age group that have received vaccinations for the infectious disease being modelled. For example, if 90% of 1 year olds in Ireland had been given the MMR vaccination in 2011 and we are running a model for 2012, we give each agent in the model with an age of 2 a 90% chance of having been vaccinated. If an agent is vaccinated they are given a 97% chance of being immune to the disease. This takes into account vaccination failure and is based on the vaccine effectiveness rate for measles (Nelson & Williams, 2007). Half of the agents with age less than 1 are given immunity to a disease to mimic passive immunity infants receive from their mothers (Nicoara et al., 1999). For any agents that have an age corresponding to a vaccination year not in our data we give a

99% chance of being immune. Prior to vaccination campaigns the majority of the population would have either had or been exposed to childhood diseases, such as measles, leaving them immune in later life.

5.2.3 Disease

In this work we study measles-like diseases and have designed a disease model to reflect this. To simulate the transmission dynamics of measles in the model we use a compartmental SEIR type model. The SEIR model categorizes agents into susceptible, exposed, infected or recovered statuses and looks at movement of agents between categories (M. J. Keeling & Rohani, 2008). Disease transmission between agents in the model occurs as follows. When an infectious agent comes into contact with an agent who is susceptible, the susceptible agent will become exposed if a random number drawn between 0 and 1 is less than the probability of infection for the disease. In the model we consider any agents who are occupying the same patch as in contact with each other. Once an agent moves to the exposed state they are assigned an exposure time which corresponds to the length of time where they will remain exposed before becoming infectious. On average people stay exposed but not infectious to measles for 10 days (Nelson & Williams, 2007). Agents are assigned an exposure time sampled from a normal distribution with mean of 10 and standard deviation of 0.5. After this time has passed the agent will become infectious. Again the agent is assigned a length of time where they will remain infectious before recovering, on average this is 8 days (Nelson & Williams,

2007). This value is sampled from a normal distribution with mean of 8 and a standard deviation of 0.5.

The probability of infection is determined using the values for R_0 for measles (12-18) (Nelson & Williams, 2007). As discussed in Section 2.1 R_0 is the basic reproductive number and is defined as the expected number of individuals infected by one infectious individual in a completely susceptible population. The parameter can be broken down into three components, number of contacts per unit time (c), the transmission probability per contact (τ), and the duration of the infectiousness (d). The relationship between these different components was presented in Equation 2.1 which is repeated here for convenience:

$$R_0 = c\tau d \tag{5.1}$$

Of the four variables in the equation, R_0, c, τ, and d, three are known or can be estimated from our model. R_0 is known to be between 12 and 18 (Nelson & Williams, 2007), d is known to be approximately 8 days or 96 time steps in our model, and we can determine c, the average number of contacts per agent per tick by running the model. Here we take c as an average across all agents, however, within the model agents movements determine their individual contact rate, thus each agent could have a different number of contacts. This number of contacts will influence the transmission of the disease or likelihood of infection, if an agent has a large number of contacts they should have a higher chance of becoming infected and will spread the disease to more agents. The model for each town is run 20

times with no infection. For each tick the agents keep track of the number of other agents they have come into contact with. This number is averaged across agents at the end of the model run and then the average is taken across the 20 runs. Once the average contacts per time step is calculated the probability of infection is calculated by rearranging the R_0 formula from Equation 2.1 as follows:

$$\tau = \frac{R_0}{cd} \tag{5.2}$$

The model was tested using values for R_0 within the range of 12-18 and the value of 12 was selected as it the result from a model with an R_0 of 12 seemed to best match the results of the 2012 measles outbreak in Schull, Ireland that's discussed later in Section 5.3.3.

Although we present the model using measles as the infectious disease it is possible to adjust the model for any airborne infectious disease where transmission is determined by SEIR dynamics, such as influenza or mumps.

5.2.4 Transportation

In the model agents use straight-line transportation. Agents will move between their home and destination in a straight line following the most direct route. Agents move in steps from one adjacent patch to the next and will reach their destination within one time step. Although this is a naive transportation model, for small towns, such as those described in this chapter, where distances travelled are short, it is a reasonable simplification.

5.2.5 Schedule

The model is run in Netlogo using discrete time steps (this is different from the equation based model presented in Chapter 3 that is a continuous time model) and runs from initialization to the point where there are no more agents who are exposed or infected. The behaviour of each agent is updated at each time step. The time steps represent two hours in a day, with 12 steps representing one day. Each time step an agent moves throughout the town. During a week day agents leave their home in the morning and go to school or work. At the end of the school/work day agents return home. Agents who are not students or working will move randomly throughout the town, choosing different locations within the community to move to during daytime hours. On weekend days all agents move randomly throughout the town during the day and during summer weeks students will move randomly throughout the town every day. Further discussion of the movement of agents can be found in the "Submodels" section of Appendix 1.

5.3 Model Evaluation

The following sections describe the evaluation and testing of the model. The evaluation is similar to that done in Section 3.3 but as there are more parameters that are involved in the agent-based model and these can be easily adjusted we run some additional tests. First we present results from tests to illustrate that the model is working as expected for a measles outbreak. We examine the infection curves created by the model, and then we run a sensitivity analysis looking at

three parameters: the probability of infection, the vaccination rate, and the chance that a student will stay home sick and self isolate. The sensitivity analysis will help to determine if the model acts as expected when parameters in the model are adjusted. For example, as vaccination rates, and thus the number of immune agents, increase the resulting size of the outbreak, and the chance that an outbreak will occur, should decrease. Similarly, if agents have a higher probability of staying home when infected, they should infect fewer agents and the outbreak should be less severe. To further test our model we run two additional experiments. The first is to show that the model is capable of recreating a historical outbreak, specifically a case study simulating a measles outbreak that occurred in Schull, Ireland in 2012. We then show how easy it is to simulate a similar outbreak in a host of towns, and, finally compare the results from the various towns. By showing that there are no obvious correlations with the model results and town characteristics we show that the model captures interactions between these characteristics and agents actions.

5.3.1 Modelling disease dynamics

When modelling an infectious disease, one important test is to determine if the model is capturing the correct disease dynamics. As described in Chapter 3 the dynamics of an SEIR model should roughly follow the curves in Figure 5.2. The SEIR infection curve plots the number of agents in each of the four categories – susceptible, exposed, infected and recovered at each time step. In order to test that our model will produce these curves, the model was run ten times for a

town (Schull, Ireland) under the case where no one in the town was vaccinated or immune. The model is run 10 times because of the inherent stochasticity in agent-based models: as a consequence of which, unlike an equation based model, one run does does not capture the full range of results for an agent-based model. Figure 5.3 shows the results of these runs in the form of the SEIR infection curve. The curves from the model match the shape of the classic SEIR curve. Fig 5.3 includes the curves for all 10 runs to illustrate that while the overall pattern of the outbreak remains the same on each run, the stochastic nature of the model means that in some runs the outbreak takes off more quickly than others. We also looked at where infections took place in the model and determined that for the town of Schull, across 100 runs, 96.5% of infections occurred in a school setting, 3% occurred at home and less than 1% occurred in work and community settings. The fact that the majority of infections occurred in a school setting indicate that school closures might be an effective intervention to stop the spread of a measles outbreak; and, in later chapters we will return to the question of interventions and school closures.

5.3.2 Sensitivity analysis

After checking the model to make sure it produced the expected disease dynamics, a sensitivity analysis was run on several parameters. The sensitivity analysis was run to determine if the model responds as expected to changes in different parameters. The parameters for infectivity of the disease, vaccination rates and

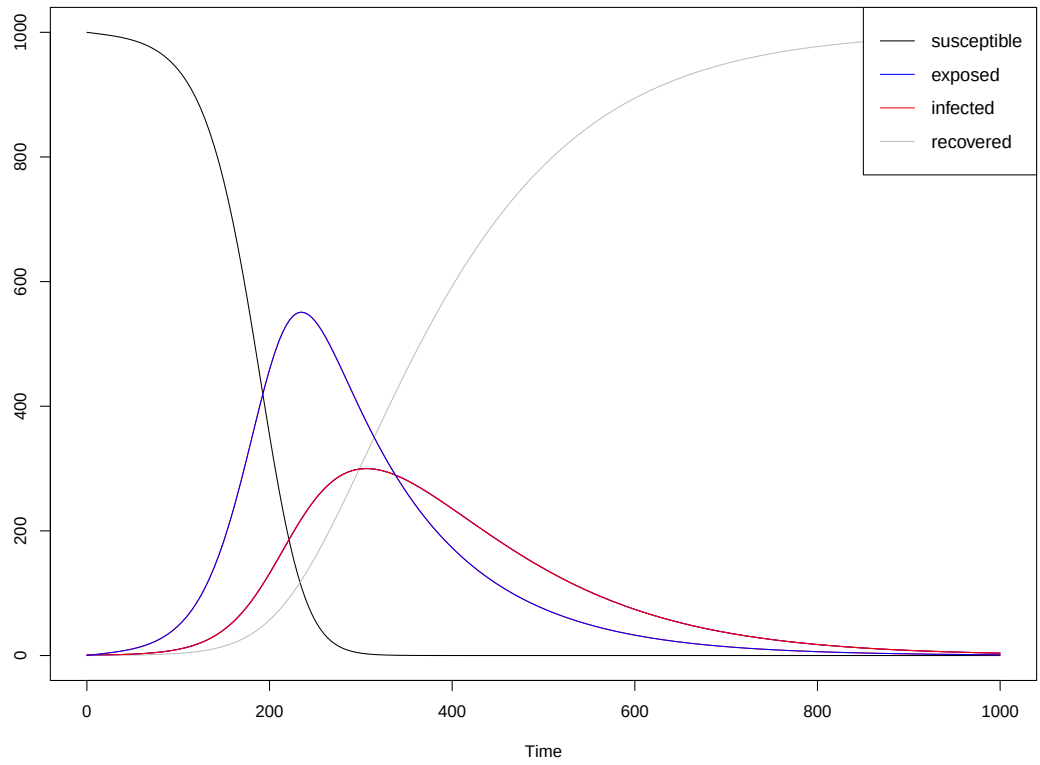


Figure 5.2: Example SEIR curves. The curves are generated using a basic SEIR differential equation model.

the percent of times an agent stays home when sick were investigated. For all analysis the models were run on the town of Schull, Ireland. The initial conditions are set so that each age category matches the census data from Schull Ireland and the proportion of agents in the vaccination categories match the all Ireland vaccination rates (unless otherwise specified). For each analyses we look at some or all of the following characteristics of the outbreak to determine how it changes: average number of agents infected, percent of outbreaks that occur and the maximum number of infected agents across runs. For the percent of outbreaks that occur we use the World Health Organization's definition of a measles outbreak to

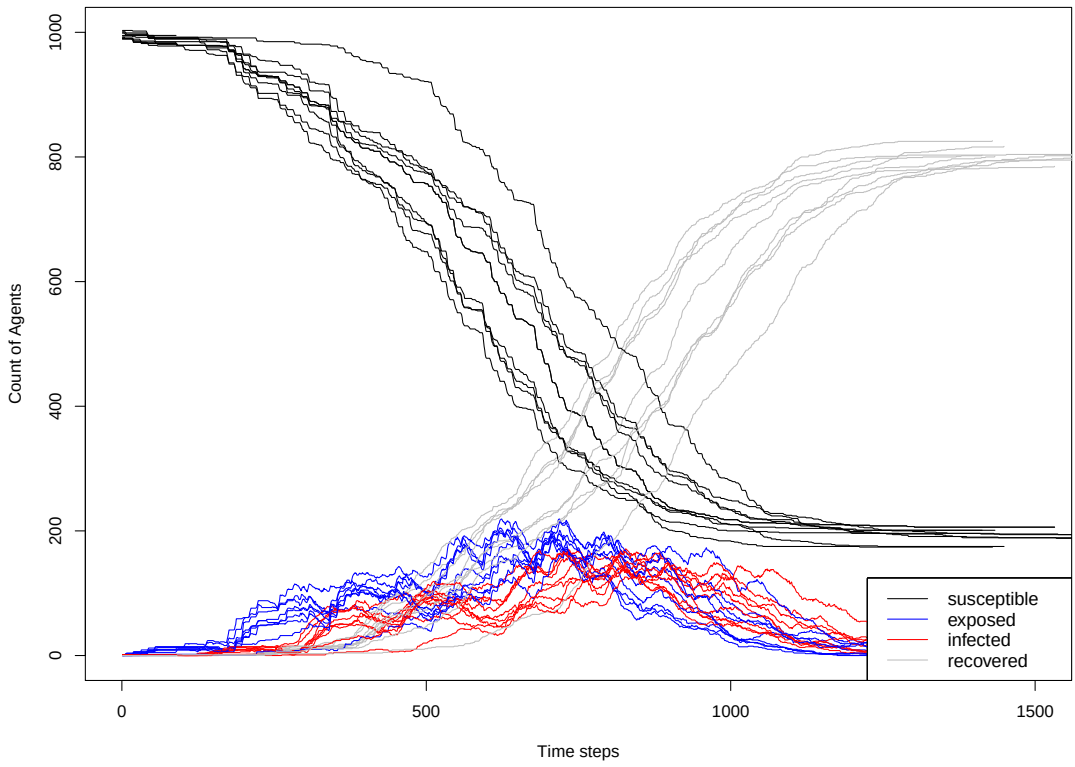


Figure 5.3: SEIR curves from the model. SEIR infection curve for 10 runs of the model in a town in which no one was vaccinated or immune.

determine when an outbreak has occurred. The World Health Organization considers a measles outbreak to be two or more cases of measles that are temporally related, and epidemiologically or virologically linked or both (*Measeles: Vaccine-Preventable Diseases Surveillance Standards*, 2018). In the model we consider an outbreak to be at least one agent infected by the initially infected agent.

Sensitivity analysis: Infectivity

To look at how changes in the probability of infection influence the resulting outbreaks we run the model 100 times each for ten different probabilities. The prob-

abilities of infections we use range from 0.01 to 0.1, the R_0 values that correspond with the probabilities of infection range from 15 to 150. These R_0 values are highly unlikely to occur in a real scenario, however, the probabilities of infection for the sensitivity analysis were not chosen because of their realistic values but instead as a method of testing the model to determine if it behaved as expected. All other parameters remain constant, including the town, which is Schull, the vaccination rates, and the home sick parameter, which is set to a 70% chance of staying home when sick. If the disease transmission is working correctly in the model there should be an increase in the size and severity of outbreaks as the probability is increased. Table 5.2 and Figure 5.4 show the results for the ten different probabilities of infection. Analysis of the results show that as the probability of infection increases, meaning that an infected agent has a higher chance of passing the virus on per contact, the average number of infected individuals, and the maximum number of infected agents all increase. There are a few times when the percent of runs resulting in an outbreak decreases by a few percentage points (from probability 0.04 to 0.05 and from 0.08 to 0.09). This result is not concerning as in both cases the decrease is only by a few percentage points and is not surprising given the stochastic elements within the simulations.

Sensitivity analysis: Vaccination rates

Three alternative vaccination scenarios were considered to determine if the disease dynamics respond as expected to changes in vaccination rates. The vaccination

Infectivity Rate	R_0	Average Number Infected	Percent Outbreaks	Max Infected
0.01	15	16.16	74	71
0.02	30	50.31	82	95
0.03	45	72.62	93	123
0.04	60	87.01	97	129
0.05	75	91.77	95	135
0.06	90	104.3	96	146
0.07	105	112.4	97	165
0.08	120	120.3	98	165
0.09	135	127.3	94	187
0.10	150	141.5	97	191

Table 5.2: Differences in model results based on changes in the probability of infection in the model.

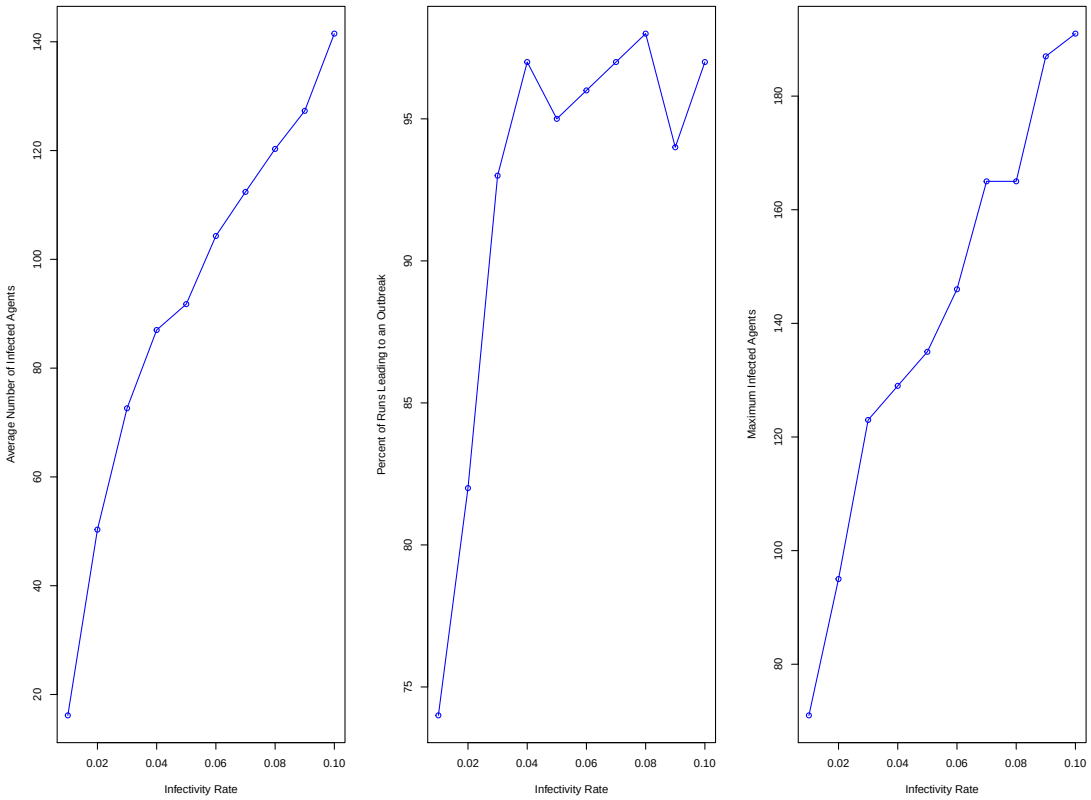


Figure 5.4: Outbreaks by probability of infection. Charts showing the change in average number of agents infected, the percent of runs leading to outbreaks and the maximum infected agents as the probability of infection changes.

scenarios used were: no vaccinations and no immunity, full vaccination and full immunity, and vaccination to the herd immunity level. All other parameters remain constant, including the town which is Schull, the probability of infection which is set to 0.008, and the home sick parameter which is set to a 70% chance of staying home when sick. If the model is working as we expect, in the full vaccination scenario we expect the outbreak should either be very small or not occur at all, in the full vaccination scenario we expect that an outbreak should not take off and in the herd immunity scenario we expect that any outbreaks that do occur should be small.

Herd immunity is the concept that there is a critical number of individuals that need to be vaccinated or immunized to interrupt the transmission of an infectious disease in a population. The concept is discussed previously in Section 2.1.

The model was run 100 times for each scenario. Results from all three scenarios are shown in Table 5.3. The average number of infected individuals across runs, the percent of outbreaks across runs and the maximum number of infected agents in any runs are compared. The main results of the analysis can be summarised as follows:

- With no agents in the town immune, outbreaks occur in 94% of runs. The average number of infected agents across the runs is 726 and the maximum number of infected agents in any run is 846. This is as expected as with no agents immune the virus can spread quickly through the town unimpeded.

- As expected with full vaccination and immunity, outbreaks do not take off.

The initial infected agent does not infect any additional agents in any run.

- In the herd immunity scenario, 52% of the runs results in outbreaks, however, the average number of infected individuals across the outbreaks is 3.28 and the maximum number of agents infected in any run is 15. The lower average infected and maximum infected numbers provide evidence that when our population reaches herd immunity the likelihood of an outbreak is significantly reduced and the size of the outbreak is also reduced.

Immunity Level	Average Infected	Percent Outbreaks	Max Infected
No Immunity	726.7	94	846
All Immune	1.0	0	1
Herd Immunity	3.3	52	15

Table 5.3: Vaccination Scenarios Sensitivity Analysis

Sensitivity analysis: Home sick parameter

The parameter to determine the percent of time an agent stays home when sick is also adjusted as another method to determine if the model is working as expected. If agents stay home more frequently when infected they will be less likely to interact with other agents and pass the infection on to others. Therefore, it is expected that if agents are more likely to stay home then the outbreak will be less severe while if an agent is less likely to stay home when infected the outbreak will be more severe. Eleven different scenarios are considered for the home sick parameter: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100%. All other parameters remain constant, including the town which is Schull, the vaccination rates, and the probability of

infection which is set to 0.008. The model is run 100 times for each new parameter value. Results can be found in Table 5.4 and Figure 5.5.

Chance of Staying Home	Average Number Infected	Percent Outbreaks	Max Infected
100%	1.0	5	2
90	2.6	35	24
80	5.4	53	35
70	15.9	74	62
60	29.3	79	97
50	42.7	88	104
40	54.8	94	103
30	64.6	97	106
20	72.8	100	108
10	79.1	100	126
0	85.1	100	121

Table 5.4: Differences in model results based on the percent chance of agents staying home when sick.

As can be seen from Table 5.4 and Figure 5.5, as the percent chance of staying home decreases the average number of infected agents and the percent of runs that lead to an outbreak increase. This is further evidence that the model is working correctly as this is what is expected.

5.3.3 Case study: Schull 2012

To evaluate the ability of our model to replicate an actual outbreak of measles we present a case study that focuses on simulating a real measles outbreak that occurred in Schull, Ireland in 2012 (HSE, September 2012). The outbreak in Schull was discussed earlier in Section 3.3.3. It is a small town with approximately 1,000 residents, roughly 30% of those residents are students and 89% are immune to measles through either having been vaccinated or from having the disease previ-

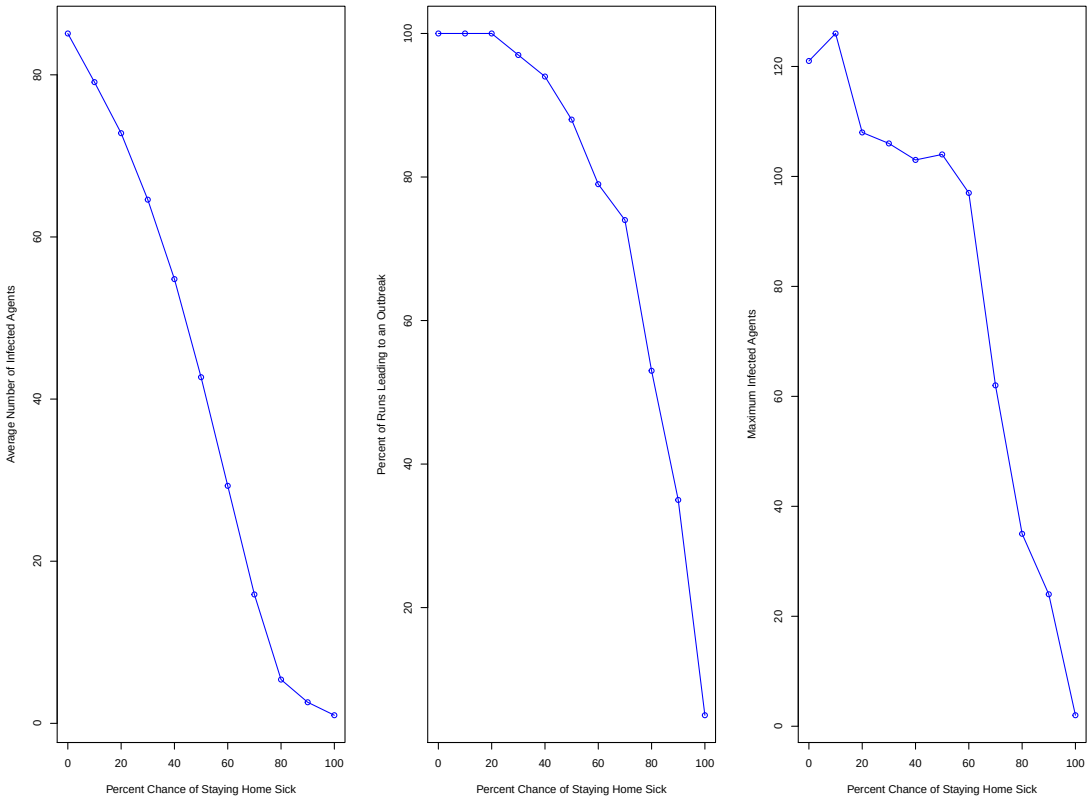


Figure 5.5: Outbreaks by percent chance of staying home sick. Charts showing the change in average number of agents infected, the percent of runs leading to outbreaks and the maximum infected agents as the percent chance of staying home sick increases.

ously.

As an agent-based model is designed to be stochastic in order to reflect the different decisions that individuals make in the real world, we do not expect that our agent-based model will identically reproduce the Schull outbreak every time it is run. In fact, we do not necessarily expect to see a perfect replication of the outbreak for a majority of runs. The Schull outbreak happened as it did because of decisions made by infected individuals: for example a decision by a person to go to school while feeling ill would most likely increase the number of infections in

the outbreak, while a decision by the same person to stay home would most likely have the opposite effect. Our agent-based model captures these different scenarios as the agents decide at each time step in the simulation what their actions are going to be.

Instead of showing that the model perfectly replicates the Schull outbreak, we want to show that the Schull outbreak falls within the range of outbreaks that our model predicts. We will show this by looking at the size of the outbreak, the ages of agents infected, the number of agents infected per week and the length of the outbreak.

In order to create a model that accurately represents the Schull outbreak, parameters need to be adjusted for the town. We select a start week of 15 which corresponds to the timing of the Schull measles outbreak starting in the 15th week of the year. We also set the probability of infection to 0.008 which is derived using Equation 2.1. West Cork vaccination rates for MMR were used and the initial infected individual is set up as a student who was not immunized. The model was run 400 times with the Schull parameters.

The stochasticity in the simulation leads to different results each time the model is run. The average number of agents infected across the runs was 17 with a maximum of 90 infected agents in one run. Twenty-five percent of the runs results in outbreaks that had more than 30 agents infected.

The results show that while the average for all the runs is lower than the number of infected in the Schull outbreak, the number of people actually infected

is in the 75th percentile of model runs. The Schull outbreak was primarily made up of individuals between the age of 10-19. The average percent of infected agents in each age group is shown in Figure 5.6.

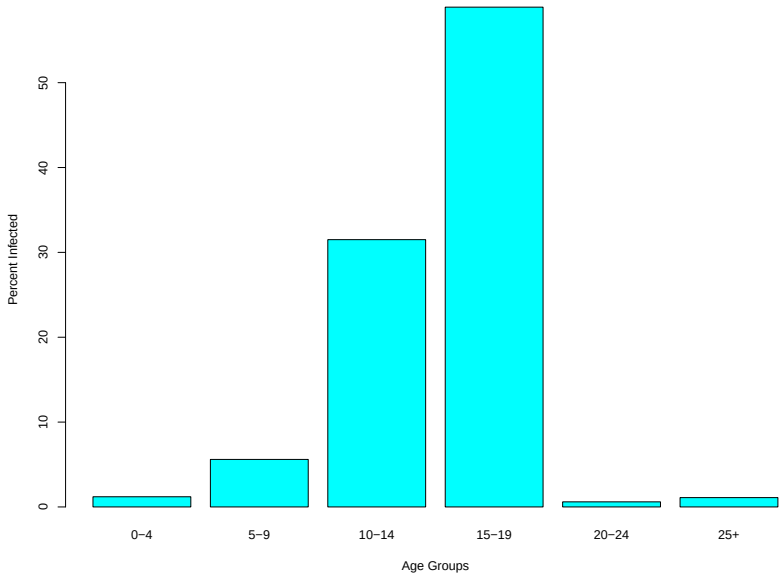


Figure 5.6: Average Number of Agents Infected in Schull by Age Groups.

The distribution of infected agents by age matches what occurred in the Schull outbreak with the majority of individuals being infected in the age groups between 10 and 19. This is taken as evidence that the model simulates a scenario similar to what actually happened.

The outbreak patterns are also analysed to determine if our model results in outbreaks that follow the same path as the actual outbreak. The reported cases for the Schull outbreak peak 3 and 5 weeks after the initial case is reported. Some of the runs match closely with the outbreak pattern for Schull and some do not. Figure 5.7 shows some of the different outbreak patterns that occur in our model.

Figures 5.7 *a* and *b* are most similar to the Schull outbreak with peaks of cases two weeks apart.

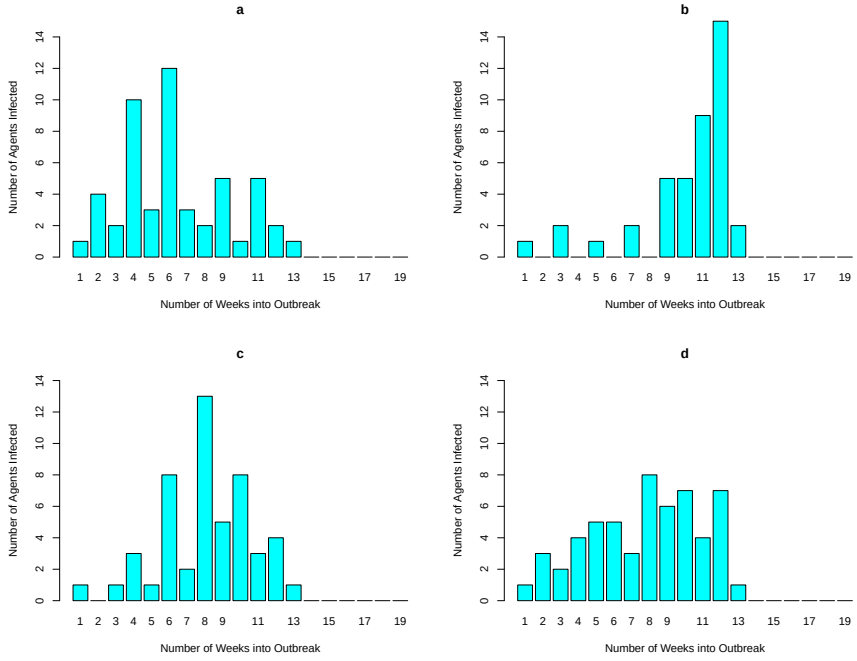


Figure 5.7: Distribution of Agents Infected by Week in Four Different Runs

Although our model does not predict perfectly what happened in Schull every run, we believe that a number of the runs do capture the outbreak scenario and that the Schull outbreak falls within the range of outbreaks in our simulation. The fact that the model produces 25% of runs with the number of agents infected similar to the Schull outbreak, the age distribution of the agents infected in the runs and that a number of runs match the distribution of cases by week found in the Schull outbreak are taken as evidence that we are capturing an outbreak similar to the one that occurred in Schull.

5.4 Simulating additional towns

The main advantages of agent-based models is their ability to capture heterogeneity (allowing each agent to potentially have a unique set of characteristics and make their own decisions) along with complex behaviours, interactions, and characteristics of agents. We add to this the ability to accurately simulate different towns based on data from publicly available data sources. We demonstrate this by simulating outbreaks similar to what happened in Schull in 32 other Irish towns. The towns are the same small to medium size towns in Ireland selected in Section 3.4. The populations of the towns are between 390 and 9,548. The areas of the towns range from 5.26 km^2 to 63.96 km^2. We show that subtle differences between these towns mean that the outbreaks follow different profiles to that seen in Schull. We pick a subset of these towns that are similar to Schull in terms of population, area, or both for deeper analysis. The following sections discuss simulating a measles outbreak on 33 different towns including Schull and the differences in those outbreaks. We then choose towns that are similar in population but not area, area but not population, or both to Schull and how the outbreaks from the simulation compare to the Schull simulation.

The towns selected are included in Table 5.5 along with a set of factors that help to define each town and the percent of model runs that lead to an outbreak for each town. The factors are as follows: the number of small areas in the town, the percent of students in the town, the percent of unvaccinated individuals in the town, and the population density. Table 5.5 also includes the effective reproductive number,

R_e, which is the reproductive number, R_0, adjusted to account for immunity in the population and was introduced in Section 2.1.

For each of the towns we simulate a measles outbreak with the same conditions as Schull. To limit the variability in the models to only characteristics of each town such as area, population size, age structure and number of schools, the vaccination rates for all of Ireland are used instead of region specific rates. Schull is rerun with the all Ireland vaccination rates in order to make the results comparable to the other towns. To calculate the transmission probability for each town we find the average contacts per time step and then use Equation 2.1. Table 5.5 has a column including the transmission probabilities for each town. All other model parameters are the same as in the Schull model from the previous section. For each town the model is run 200 times and the percent of runs where an outbreak occurs (the initially infected agent passed the virus to at least one other agent) is determined. The percent of runs with outbreaks (two or more cases of measles) for each town is listed in the rightmost column in Table 5.5. We can see that the number of runs that result in an outbreak in each town are quite different. Even in some cases where the size and populations of the towns are similar (e.g. Bagenalstown and Ballyjamesduff) we see quite different outcomes. This is evidence of the value of agent-based modelling and shows the value in being able to simulate towns accurately based on publicly available data sources.

To get a better understanding of the relationship between factors and outbreaks, Figure 5.8 shows a scatter plot matrix of the seven factors and the percent

of runs that lead to an outbreak. To examine the relationships further we calculate the Pearson correlations between each factor. When analyzing the correlations we use the following guidelines for interpreting the coefficients: 0 corresponds to no linear relationship, 0 to 0.3 or 0 to -0.3 corresponds to a weak linear relationship, 0.3 to 0.7 or -0.3 to -0.7 corresponds to a moderate relationship and 0.7 to 1.0 or -0.7 to -1.0 corresponds to a strong linear relationship (Ratner, 2009). The correlations for our model are presented in Table 5.6. From the scatter plots it can be determined that none of the factors have a clear relationship with the percent of runs that lead to an outbreak. This is further shown in the correlation table, it can be seen that there is no strong correlation between percent of runs that result in an outbreak and any of the factors. This analysis shows that no single characteristic of a town overly impacts the likelihood of an outbreak, but rather that this is governed by the heterogeneity allowed by an agent-based model and the interactions between agents that are simulated. We can see some relationships between various factors. Population and small areas have a strong positive linear relationship, with a correlation of 0.970. This is as expected as small areas are defined as geographic regions with between 50 to 200 households. A larger population will have more households and thus more small areas. Additionally the percent of unvaccinated agents in the town appears to have a moderate linear relationship with both the percent of students in the town and the population density. The correlation between the percent of unvaccinated agents and students is 0.589 and between the percent of unvaccinated and the density is 0.515. This also makes sense, as the

measles vaccination was introduced in 1985 in Ireland. Thus the older population is largely immune due to contracting measles while the younger population would be the ones getting vaccinated. With the same vaccination rates by age across towns, a town with higher percentage of students should have a higher percentage of unvaccinated individuals across the whole town.

Town	Population	Area(km^2)	Probability of Infection	Small Areas	Percent Students	Percent Not Immune	Density	Outbreak	Confidence Interval	R_e
Arainn	1,251	47.48	0.009	6	30.7	14.7	26.34	69.5	(63.1, 75.9)	1.76
Ardamine	3,293	23.33	0.006	19	33.8	12.4	130.00	88.5	(84.1, 92.9)	1.49
Ardfert	997	7.97	0.006	4	38.5	13.4	125.09	65.0	(58.4, 71.6)	1.61
Arranmore	514	18.08	0.055	4	31.3	8.4	28.43	69.5	(63.1, 75.9)	1.01
Bagenalstown	3,421	18.00	0.003	13	35.2	12.8	190.06	66.0	(59.4, 72.6)	1.54
Ballyjamesduff	3,134	21.60	0.005	12	37.4	13.5	145.09	83.5	(78.4, 88.6)	1.62
Banagher	1,993	19.85	0.009	8	36.5	12.7	100.40	88.5	(84.1, 92.9)	1.52
Blarney	5,310	23.30	0.001	21	36.9	13.2	227.90	60.0	(53.2, 66.8)	1.58
Castlereagh	3,077	40.09	0.013	15	26.3	10.2	76.75	85.5	(80.6, 90.4)	1.22
Clane	7,527	18.89	0.002	28	36.5	14.5	398.46	86.0	(81.2, 90.8)	1.74
Croom	1,690	18.17	0.004	6	35.7	11.4	93.01	57.0	(50.1, 63.9)	1.37
Donegal	4,010	31.49	0.006	17	30.4	12.5	127.34	92.0	(88.2, 95.8)	1.5
Gort	2,671	11.21	0.009	12	27.6	11.6	238.27	87.5	(82.9, 92.1)	1.39
Kenmare	2,912	55.61	0.006	17	29.4	10.3	52.36	81.5	(76.1, 86.9)	1.24
Kilcock	6,234	16.40	0.001	23	35.7	14.4	380.12	58.0	(51.2, 64.8)	1.73
Kildare	9,325	37.09	0.002	32	36.1	14.0	251.42	88.5	(84.1, 92.9)	1.68
Kilkee	1,037	5.26	0.008	8	25.7	9.1	187.15	65.5	(58.9, 72.1)	1.10
Killadysert	922	63.96	0.009	4	36.1	10.0	14.52	61.0	(54.2, 67.8)	1.20
Kinsale	6,871	12.96	0.003	31	30.4	11.5	129.03	87.0	(82.3, 91.7)	1.38
Lisdoonvarna	861	12.96	0.010	3	26.5	12.3	66.44	63.0	(56.3, 69.7)	1.48
Louisburgh	983	23.30	0.009	7	27.6	11.8	42.19	62.5	(55.8, 69.2)	1.42
Moate	3,046	21.34	0.007	12	32.6	12.7	142.74	90.0	(85.8, 94.1)	1.52
Oranmore	4,325	22.38	0.002	18	28.9	13.1	193.25	62.0	(55.3, 68.7)	1.57
Portmagee	390	16.77	0.023	2	27.7	10.3	23.26	61.5	(54.8, 68.2)	1.24
Rathnew	3,294	6.90	0.003	10	37.4	15.0	477.39	73.0	(66.8, 79.2)	1.80
Roscrea	6,318	48.45	0.006	26	33.2	11.5	130.40	91.0	(87.0, 96.0)	1.38
Rosslare	2,057	17.90	0.003	12	26.0	9.1	114.92	47.0	(40.1, 53.9)	1.10
Roundstone	459	28.01	0.041	4	31.1	11.1	16.39	86.0	(81.2, 90.1)	1.33
Schull	987	17.03	0.008	7	30.6	11.4	57.96	72.5	(66.3, 78.7)	1.37
Shanagolden	946	17.79	0.008	4	33.0	11.0	53.18	54.5	(47.6, 61.4)	1.32
Stamullin	4,694	37.68	0.003	14	37.0	12.9	124.58	78.5	(72.8, 84.2)	1.55
Strokestown	1,003	18.11	0.009	6	31.4	11.1	55.38	73.5	(67.4, 79.6)	1.33
Tramore	9,548	16.60	0.001	36	35.8	12.2	575.18	73.0	(66.8, 79.2)	1.46

Table 5.5: Area, population and other characteristics for each of the 33 selected towns

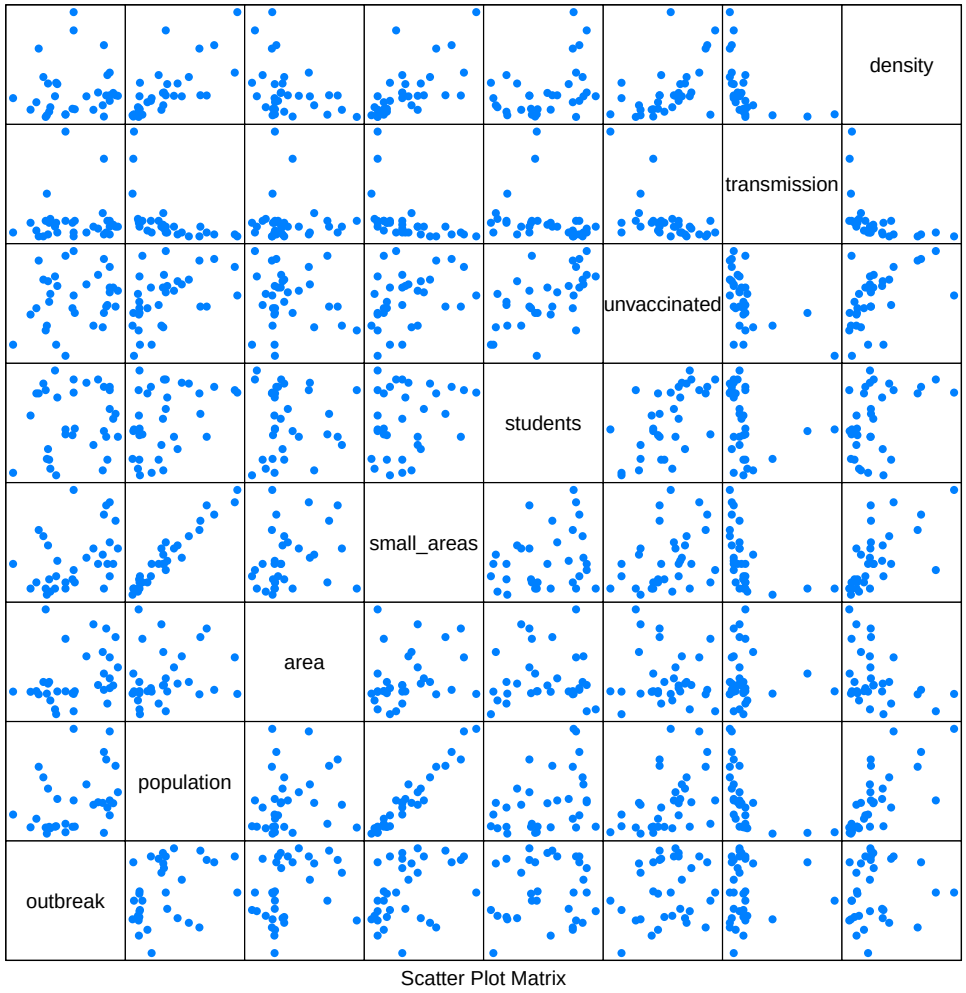


Figure 5.8: Percent of Runs Leading to an Outbreak. Scatter plot of percent of runs resulting in outbreak and factors defining each town. *Outbreak* is percent of runs resulting in outbreaks, *small_areas* is the number of small areas in the town, *students* is the percent of students in the town, *unvaccinated* is the percent of unvaccinated agents in the town, *density* is the population density, and *transmission* is the probability of transmission per contact.

	Outbreak	Population	Area	Small Areas	Students	Unvaccinated	Transmission	Density
Outbreak	1	0.362	0.312	0.399	0.112	0.208	0.044	0.077
Population	0.362	1	0.203	0.970	0.388	0.462	-0.483	0.718
Area	0.312	0.203	1	0.263	0.001	-0.105	-0.048	-0.347
Small Areas	0.399	0.970	0.263	1	0.251	0.340	-0.462	0.648
Students	0.112	0.388	0.001	0.251	1	0.589	-0.260	0.407
Unvaccinated	0.208	0.462	-0.105	0.340	0.589	1	-0.509	0.515
Transmission	0.044	-0.483	-0.048	-0.462	-0.260	-0.509	1	-0.440
Density	0.077	0.718	-0.347	0.648	0.407	0.515	-0.440	1

Table 5.6: Correlation table for percent outbreaks and the other town characteristics

5.4.1 Towns similar to Schull

We further break our analysis down to towns that have characteristics similar to Schull. Intuitively it would make sense that an infectious disease outbreak in two towns of approximately equal population and area would be similar. However, we hypothesize that this is not always the case and that interactions between both known characteristics that are programmed into the model and other more intangible characteristics that emerge from the model lead to differences in outbreaks. In order to test this, we select twelve of the towns from the previous analysis that have similar population sizes, town area, or both to Schull. Two towns selected are similar in both area and population, four are similar in population and six are similar in area. Table 5.7 gives the areas, populations, transmission probability and percent of runs that result in an outbreak for the twelve towns plus Schull.

A Kruskal-Wallis test is done to compare the results across the towns. The test is a non-parametric test to determine if sample distributions comes from the same population distribution. In this case if the distributions of the number of agents infected across the 200 runs for the different towns could be from the same population. The null hypothesis of the test is that the samples are from the same

Town	Percent Outbreaks	Population	Area	Transmission
Ardfert	62.5	997	7.97	0.006
Bagenalstown	65.6	3,421	18.00	0.003
Croom	57.5	1,690	18.17	0.004
Kilcock	57.25	6,234	16.40	0.001
Kilkee	66.5	1,037	5.26	0.008
Killadysert	59.3	922	63.96	0.009
Louisburgh	67.5	983	23.30	0.008
Portmagee	57.5	390	16.77	0.023
Rosslare	50.5	2,057	17.90	0.003
Schull	72.5	987	17.03	0.008
Shanagolden	58.3	946	17.79	0.008
Strokestown	74.0	1,003	18.11	0.009
Tramore	73.0	9,548	16.60	0.001

Table 5.7: Percent outbreaks, area and population for each of the 12 selected towns and Schull

distribution. The test results in a p-value of <0.0001. This results in a rejection of the null hypothesis leading to the conclusion that there are statistical differences between the distributions of outbreaks in the towns. To get a better idea of how area and population affect the outbreaks, towns with area similar to Schull are analysed separately from towns with population similar to Schull and towns with both population and area similar to Schull.

Population

The four towns selected with a similar population to Schull but a different area were Ardfert, Kilkee, Killadysert, and Louisburgh. The percent of runs resulting in outbreaks for the four towns range from 59.3% for Killadysert to 67.5% for Louisburgh. All of which are lower than the percent of runs resulting in outbreaks from the Schull model. Running a Kruskal-Wallis test across the percent of runs

resulting in outbreaks for each of the four towns and Schull gives a p-value of 0.001 resulting in a rejection of the null hypothesis that the samples (in this case the percentage or runs resulting in outbreaks for each town) are from the same distribution. Therefore, it can be concluded that other factors besides population size influence the course of an outbreak.

Area

Six towns were selected because they had area similar to Schull. The towns are Croom, Portmagee, Tramore, Bagenalstown, Rosslare and Kilcock. The percent of runs resulting in outbreaks for the six towns range from 50.5% for Rosslare to 73% for Tramore. The percent of runs resulting in outbreaks from the Schull model is in this range, with 72.5% of runs for Schull with more than one agent infected. Running a Kruskal-Wallis test on the percentage of runs resulting in an outbreak for the six towns plus Schull gives a p-value of <0.0001 resulting in the rejection of the null hypothesis, which leads to the conclusion that the area of the town does not determine the percentage of runs that lead to an outbreak.

Area and population

Strokestown and Shanagolden were selected as two towns that shared many characteristics to Schull. Both towns have similar area and population size but additionally have similar age structures. Looking at the percent of students in the population, the group believed to be most susceptible to an outbreak of measles, Schull has 31% students in the town, Strokestown has 31% students and Shanagolden has

33% students. Additionally the percent of individuals in each town who are not vaccinated and thus susceptible to the infection is compared. Although vaccination rates are constant across towns in the model because the vaccination rates are age specific the overall percent of vaccinated individuals in a town may vary due to different age structures. Schull, Strokestown and Shanagolden all have 11% of the population who are not vaccinated or not immune.

Comparing the results for the three towns, Schull and Strokestown both have similar percents of outbreak runs, with 72.5% of runs for Schull with more than one agent infected and 74% of runs for Strokestown with more than one agent infected. Shanagolden, however, has different results with 58.3% of runs with more than one agent infected. The differences can be seen in more detail by looking at the overall distribution of agents infected by run for the three towns. Figure 5.9 shows histograms showing the percent of runs by number of agents infected for the three towns.

Although Schull and Strokestown have similar results, Shanagolden's lower proportions of outbreaks illustrate that within our model similar towns can have different results. This is to be expected as it is likely that interactions between characteristics and the town layout lead to different outbreaks. These results also emphasize why an agent-based model is important in looking at infectious disease outbreaks. Other modelling methods such as equation based models would not capture these interactions. Although we are not able to determine the exact reason for differences between towns we can make some guesses. One possible difference

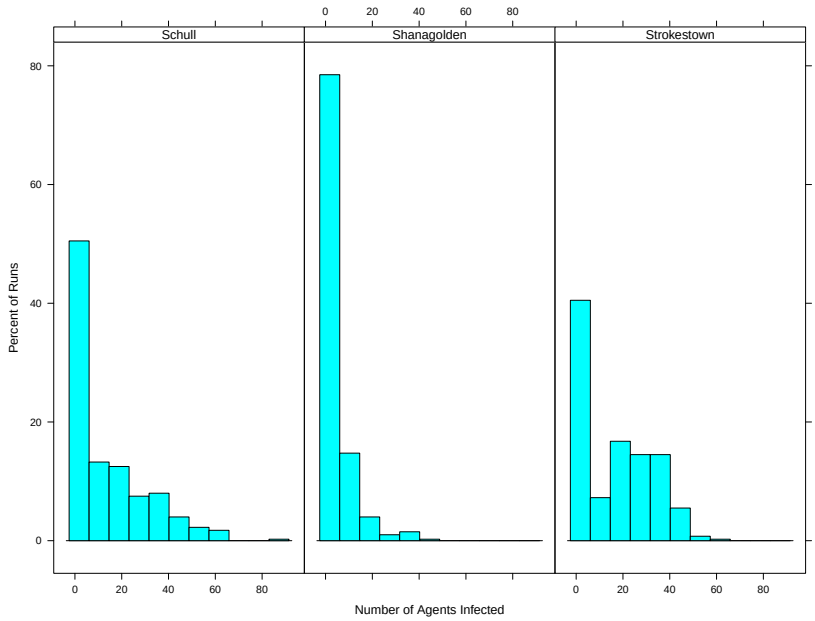


Figure 5.9: Histograms showing the percent of runs by number of agents infected for Schull, Shanagolden, and Strokestown

between the towns is the number of schools. Strokestown has two primary schools while the other two towns have only one. Another possible interaction that could lead to differences has to do with the town layout. Both Strokestown and Schull appear to have a larger town center while Shanagolden appears more spread out. Figures 5.10, 5.11 and 5.12 show the small areas of the towns color coded based on population density. The darker the green color the higher population density. From the figures one can see that both Schull and Strokestown have a few higher density small areas compared to Shanagolden. Both these factors could lead to differences in agent interactions and thus model results.

Based on the results we can conclude that the agent-based model is capturing interactions between the factors and the agents actions and these interactions are what is defining the outbreak. If it was only factors that were programmed into

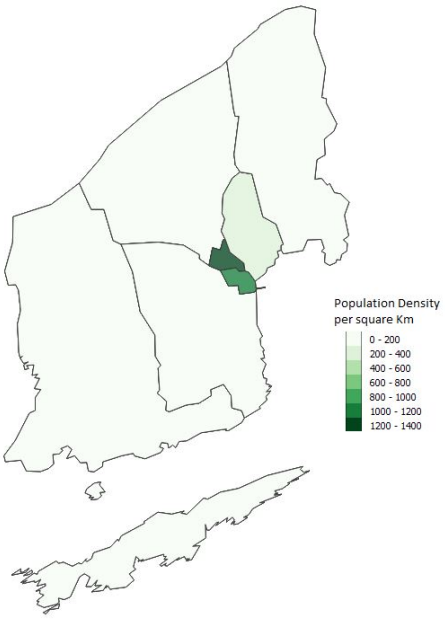


Figure 5.10: Schull, Ireland Map showing the population density per sqkm in Schull from the 2011 Census CSO (2014a).

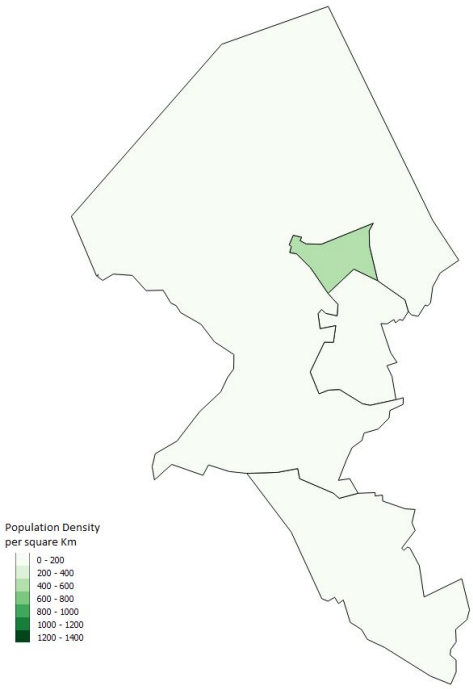


Figure 5.11: Shanagolden, Ireland Map showing the population density per sqkm in Shanagolden from the 2011 Census CSO (2014a).

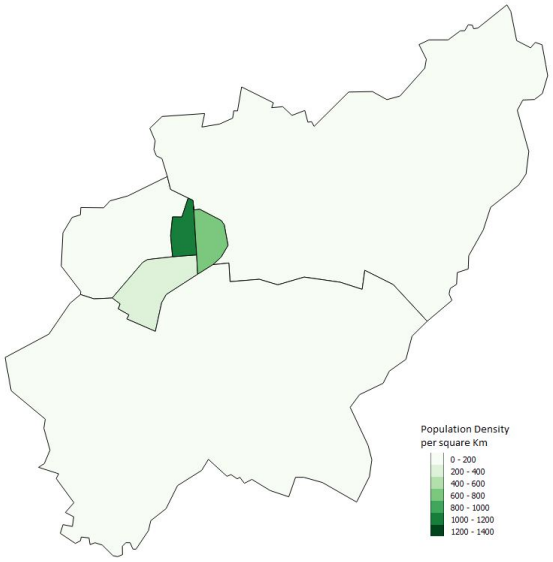


Figure 5.12: Strokestown, Ireland Map showing the population density per sqkm in Strokestown from the 2011 Census CSO (2014a).

the initial setup of the model that affected the course of the outbreak we would expect a stronger relationship between the percent of runs resulting in an outbreak and the various factors that we discuss.

5.5 Conclusion

We have presented an approach to creating an open data-driven agent-based model for human infectious disease epidemiology. By simply changing a few parameters we are able to quickly and easily adjust the model to simulate an outbreak in any town in Ireland as long as the data for the town is accessible. Similarly if the data is available we are confident that our model can simulate the spread of measles in a non-Irish town. Although we have only presented models for the spread of measles in this thesis, with the adjustment of disease modelling parameters,

such as probability of infection and exposure time, the model can be adjusted to simulate the spread of any infectious disease that follows the SEIR person to person transmission dynamics.

Using an agent-based model allows one to capture the stochasticity that exists in a real world system. An agent-based model does not just give one prediction but a range of possible outbreaks that may occur as agents are allowed to make decisions similar to how individuals in the real world will make decisions. Running the agent-based model multiple times will capture different possible scenarios for the outbreak that are all determined by how the agents interact. For example, when simulating the Schull outbreak, model runs where the outbreak is less severe than the Schull outbreak could be cases where individuals stayed home when infected, and, therefore, did not infect other individuals. Runs where the outbreak is more severe could be due to the lack of inclusion of prevention measures in our model. This is similar to what could happen in a real world scenario, the actions of a few individuals could alter the course of an outbreak, and is important when considering the effects of future outbreaks.

We believe that such a model could eventually be useful as a public health tool. Understanding how different interventions may affect an outbreak or how resilient a population already is can have a major influence on prevention strategies and preparation. Being able to easily model similar outbreaks in different towns can be an important tool in resource allocation, for example deciding where to focus a vaccination campaign when the resources do not exist to focus on every town.

However, the model presented in this chapter only simulates the spread of measles in one town. Although this can be useful in determining susceptibility of a given town, other factors not included in a closed town model such as commuting and the centrality of the town within a network of other towns can have an large influence on outbreaks and should be considered when creating a model. We will return to the questions of integrating commuting and town centrality in a network later in the thesis. In the next chapter, however, we will focus on comparing the Equation Based Model from Chapter 3 with this basic Agent-Based Model.

Chapter 6

Comparison of Agent-Based and Equation Based Models

Chapters 3 and 5 give results for our equation based model and agent-based model for the spread of an infectious disease through an Irish town. It is important to carefully look at the results of both models. If the easier to implement equation based model does just as good of a job in simulating an outbreak as the agent-based model then the added complexity of the agent-based model does not add to the results of the model. Thus to go with an agent-based model over an equation based model it is necessary to make sure that the advantages of the model outweigh the additional costs in running the model. We start our comparison by comparing the results of the equation based and agent-based model for 33 different towns in Section 6.1 and then we look at how both agent-based models and respond to introducing an intervention by changing vaccination rates in the model in Section

6.2.

6.1 Model Results

To compare the two models we have simulated a measles outbreak in the same 33 towns. For the purpose of comparison we take as the outcome of an outbreak[1] the final recovered number from the equation based model, this is the number of individuals in the model who started as susceptible, became exposed, moved to infected and then recovered and represents the magnitude of the outbreak. For the agent-based model we use the average number of infected agents across 200 runs, as agent-based models have stochasticity in the model each run can have different results. For each model run there is a total number of immune agents, these are the agents who have been infected and recovered. We take the average across the 200 runs for each town and find an average magnitude of the outbreak for the town. The agent-based model provides us with a range of outbreaks that might occur in the current system. The average does not need to be calculated for the equation based model because there is no stochasticity in the system. The lack of stochasticity in the model means that the same result is found each time. Table 6.1 presents the results for the equation based model for each town, the average number of immune agents from the agent-based model including the confidence intervals and the maximum number of infected agents across all runs of the agent-based model. Looking at the results it can be noted that in some

[1]We define an outbreak as two or more cases of measles. This is based off of the WHO's definition of a measles outbreak

towns such as Arainn and Schull the results are similar between the agent-based model and the equation-based model. However, many of the towns have different results between the two models, showing that the results can vary and that a comparison is necessary. It is important to note that there are only two towns, Blarney and Kilcock, where the number of infected people in the equation based model is greater than the maximum number infected in the agent-based model. This means that for all other towns the outbreaks produced from the agent-based model include outbreaks of the same size as the outbreaks produced from the equation-based model.

To determine which characteristics of the towns (e.g. population and area) are most related to the results of the model simulations we calculated Pearson correlation coefficients between simulation results and town descriptors. Table 6.2 shows the correlations between the results of the agent-based and the equation based model and other variables including the population size, percent of unvaccinated individuals, percent students in the town and the area of the town. From the correlations it can be seen that the results for both models are correlated with the population of the town. However, the equation based model has a much higher correlation with population than the agent based model.

In fact the agent-based model has a higher correlation with the equation based model results than with the population. From the correlations it can also be seen that the equation based model is correlated with the percent of unvaccinated individuals and the percent of students in the town. This makes sense as these are

variables that are included in the model. Area, however, has a near zero correlation with the equation based model as the variable is not included. The agent-based model has a small correlation with the percent of unvaccinated agents, the percent of students in the town and the area. The smaller correlations compared to the equation based model are believed to be because the results of the agent-based model are not simply determined by the variables programmed into the model but by the individual agent decisions and interactions between the variables that are not found within the equation based model. The correlations help to show that agent-based model captures more complex relationships and interactions between variables than the equation based model.

6.2 Experiment and Results

In this section we look at how both the equation based model and the agent-based model respond to introducing an intervention that occurs before the start of the outbreak such as a change in vaccination policies. This would change the initial conditions of the model. The vaccination policy we consider is for all non school age individuals we leave the vaccination levels the same, however, for any individual in the model who is in school we implement the policy that all school age children must be vaccinated for measles. This is a policy that is implemented in many states in the USA, and France, and has been discussed as a potential policy for Ireland (Clarke, 2018). To account for children with medical exemptions we use the average percent of children in the USA that do not receive the MMR

vaccination due to medical reasons, resulting in a 99.25% vaccination rate among school children in the model (Seither et al., 2014).

We run the equation based model and the agent-based model with the new vaccination policy for the town of Kinsale, Ireland. We picked Kinsale because it was determined, based on a detailed analysis of the Irish census data, that Kinsale was the town that statistically best represents Ireland (Newstalk, 2017).

The change in initial conditions, representing the new vaccination policy, in the equation based model for Kinsale results in a reduction of the size of the outbreak from an outbreak of size 23 individuals to an outbreak of size 2. When the initial conditions are adjusted in the agent-based model there is a reduction in the average number of infected agents from 438 to 129, a reduction in the maximum number of infected agents from 655 to 534 and a reduction in the percent of runs that lead to an outbreak from 87% to 63%.

Both models show that a reduction in the outbreak size occurs when changing the vaccination policy, however, because we have a range of outbreaks that might occur with the agent-based model results we are better able to understand how the vaccination policy might influence an outbreak in the real world. It is highly unlikely that a real outbreak will match the equation based model results exactly. It is much more likely that a real outbreak will fall into the range of our agent-based model results. Thus we can show a vaccination policy will influence the likelihood of an outbreak occurring along with how it will reduce the magnitude of an outbreak if it does occur.

Town	Population	Area (km^2)	Population Density	Average ABM	Confidence Interval	Max Infected ABM	Total Infected Mathematical Model
Arainn	1,251	47.48	26.35	11.64	(9.91, 13.38)	62	10.65
Ardamine	3,293	23.33	141.15	173.96	(160.74, 187.18)	334	44.80
Ardfert	997	7.97	125.09	7.77	(6.50, 9.02)	47	24.60
Arranmore	514	18.08	28.43	15.95	(14.25, 17.64)	59	5.40
Bagenalstown	3,421	18.00	190.06	10.16	(8.78, 11.53)	79	52.80
Ballyjamesduff	3,134	21.60	145.09	219.27	(206.90, 231.63)	406	106.33
Banagher	1,993	19.85	100.40	118.17	(112.66, 123.68)	220	38.05
Blarney	5,310	23.30	227.90	5.08	(4.71, 6.30)	52	82.27
Castlereagh	3,077	40.09	76.75	170.82	(159.66, 181.97)	277	14.90
Clane	7,527	18.89	398.46	687.12	(639.60, 734.63)	1024	219.75
Croom	1,690	18.17	93.01	5.80	(4.91, 6.69)	88	62.33
Donegal	4,010	31.49	127.34	266.22	(256.37, 276.97)	415	35.97
Gort	2,671	11.21	238.27	185.99	(177.30, 196.68)	340	38.17
Kenmare	2,912	55.61	52.36	69.07	(60.91, 77.23)	234	14.81
Kilcock	6,234	16.40	385.61	6.26	(5.36, 7.16)	56	162.07
Kildare	9,325	37.09	251.42	871.69	(816.38, 926.98)	1206	259.71
Kilkee	1,037	5.26	197.15	10.05	(8.89, 11.22)	62	8.16
Killadysert	922	63.96	14.42	8.08	(7.04, 9.13)	62	9.61
Kinsale	6,871	12.96	530.17	438.14	(410.02, 466.26)	655	22.92
Lisdoonvarna	861	12.96	66.44	7.21	(6.38, 8.04)	45	12.08
Louisburgh	983	23.30	42.19	8.70	(7.23, 9.67)	51	11.50
Moate	3,046	21.34	149.75	258.04	(243.56, 274.51)	401	55.86
Oranmore	4,325	22.38	193.25	8.29	(6.30, 10.27)	114	90.72
Portmagee	390	16.77	23.26	4.67	(4.09, 5.25)	40	6.17
Rathnew	3,294	6.90	477.39	25.28	(22.29, 28.26)	176	85.89
Roscrea	6,318	48.45	130.40	638.40	(620.19, 656.60)	843	146.30
Rosslare	2,057	17.90	114.91	3.60	(3.11,4.10)	40	11.01
Roundstone	459	28.01	16.39	21.43	(19.36, 23.49)	48	15.57
Schull	987	17.03	57.96	13.88	(12.33, 15.44)	88	12.57
Shanagolden	946	17.79	53.18	4.73	(4.09, 5.36)	42	7.26
Stamullin	4,694	37.68	124.58	173.70	(150.35, 197.05)	516	95.05
Strokestown	1,003	18.11	55.38	16.86	(15.40, 18.33)	63	13.14
Tramore	9,548	16.60	575.18	44.63	(37.35, 51.91)	228	141.75

Table 6.1: Area, population and model results for each of the 33 selected towns

	EBM	ABM	Population	Unvaccinated	Students	Area
EBM	1	0.700	0.844	0.596	0.543	0.005
ABM	0.700	1	0.674	0.286	0.209	0.333
Population	0.844	0.674	1	0.462	0.388	0.203
Unvaccinated	0.596	0.286	0.462	1	0.589	-0.105
Students	0.542	0.209	0.388	0.589	1	0.001
Area	0.005	0.333	0.203	-0.105	0.001	1

Table 6.2: Correlation table for size of the outbreaks and the other town characteristics

6.3 Conclusion

Through analysis of the results of the comparison between agent-based models and equation based models we show that the two models can appear to produce very different outbreaks, however, in most cases the equation based model is in the range of outbreaks that the agent-based model produces. Looking at the correlations between the results of the models and factors that are coded into or included in the model we can conclude that the agent-based model captures interactions that the equation based model does not.

As the agent-based model follows agents through the environment it can provide us with more detailed information such as where an agent becomes infected or who infected them. This could lead to a better understanding of how the disease spreads and allow public health officials to focus on specific areas. While some exploratory work, such as looking at overall changes in vaccination rates, can be done easily in both agent-based and equation based models, something such as including push vaccinations during an outbreak or studying the effects of isolation are more complicated when done with an equation based model requiring additional equations and interaction terms, while the same could be done in an agent-based model with the introduction of a few extra behavioural rules. It

should be noted, however, that one of the advantages of the agent-based model comes in its adaptability. In order to add push vaccinations or change contact patterns the same agent-based model could be used just with different parameters while a new equation based model would need to be created.

Using an agent-based model allows one to capture the stochasticity that exists in a real world system. Agents are allowed to make decisions similar to how individuals in the real world will. Running the agent-based multiple times will capture different possible scenarios for the outbreak that are all determined by how the agents interact. For example, if the initial infected agent decides to self isolate and does not come into contact with other individuals once they know they are sick the outbreak will be much smaller than if the agent does not stay home at all. This is similar to what could happen in a real world scenario.

The equation based model does not capture these different decisions and simply presents one course of the outbreak. The equation based model, however, does have the advantage of time and computing power. Running 200 runs of the agent-based model depending on the population size and area can take days while the equation based model takes seconds. Despite the extra time it takes to run the agent-based model we feel that the results show that it has more advantages over the equation based model when trying to capture the true course of an outbreak.

Based on the comparison we feel it is important to focus on the agent-based model as the main component of our infectious disease model as we think that the advantages of agent-based models outweigh the disadvantages and provide more

information about possible outbreaks than an equation based model.

Chapter 7

Methodology: Evaluating, Validating and Testing

As discussed in Chapter 4 there is no set method within the literature for creating an agent-based model for infectious diseases. Without a standard methodology it can not only be difficult to create a model but also to show that the model has been properly evaluated and tested. The taxonomy we present in Chapter 4 creates a standard way to classify existing models and a tool to help determine what components are necessary in creating a new agent-based model for infectious diseases. However, once a model is created there are a number of steps that need to be done in order to show that the model is performing as it should and can make accurate predictions. The inherent stochasticity of agent-based models needs to be taken into account when considering the results of the model, and the model needs to be validated and tested. In Sections 5.3 and 5.4 we already began to evaluate

and test our model, however, this chapter formalizes the methodology used in those sections and generalizes it so it can be applied to future models. In Section 7.1 we discuss a method to determine the appropriate number of runs needed to account for stochasticity. Section 7.2 explains the methods used to validate the model and in Section 7.3 we present the final step in the methodology, model testing.

7.1 Stochasticity and Confidence Intervals: Selection of Number of Runs

One question that can often arise in an agent-based modelling project is how many times the model should be run. Too few times and the results might not be an accurate depiction of the model and too many times and potentially a large amount of computing and processing time has been wasted. There is no standard way to choose the number of runs necessary in the literature. This may be because of the large range of agent-based models in existence with no real modelling standards within or across fields. In Chapter 5 we run our models 200 times for each set of initial conditions (each different town). However, the question remains was that enough runs to accurately account for the stochasticity in the model or did we waste time in running the model 200 times. To determine an ideal number of simulation runs we propose using the size of the confidence interval combined with the stability of the results. Our goal is to have a relatively small confidence interval for our output statistic and to reach a point where the output is not

changing drastically with additional runs.

7.1.1 Experiment

To find the number of runs necessary to find a stable result in our agent-based model we selected three towns modelled in Chapter 5 with varying characteristics. The towns are Schull, Tramore and Kenmare and the populations and areas for these towns can be found in Table 7.1. For each town we run the model 1000 times keeping all factors besides those related to the town population, such as age specific vaccination rates, area or town layout, constant.

Town	Population	Area (km^2)
Kenmare	2912	55.61
Schull	987	17.03
Tramore	9548	16.60

Table 7.1: 2011 population and town areas for the three towns

Once the runs were completed we look at one main output statistic, the percent of runs that lead to an outbreak. The results for each town after 1,000 runs can be found in Table 7.2

Town	Percent	Confidence Interval
Kenmare	77.4	(74.8 80.0)
Schull	66.6	(63.7 69.5)
Tramore	70.0	(67.2 72.8)

Table 7.2: Percent of runs leading to an outbreak and confidence intervals by town for 1,000 runs.

To look closer at the number of runs that are actually needed for our model we take 200 samples of different sizes from the 1,000 runs. The samples are chosen as follows: the first sample contains the first 5 runs, the second sample contains the

first 10 runs, the third sample contains the first 15 runs, until the 200th sample which contains all 1,000 runs. For each sample we look at the the percent of runs that lead to an outbreak and the size of the confidence interval for the percent of runs that lead to an outbreak. Table 7.3 shows selected values of the statistics and confidence intervals for different sample sizes.

Town	Sample Size	Percent	Confidence Interval
Kenmare	5	80.0	(44.9 115.1)
	50	80.0	(68.9 91.1)
	300	77.6	(72.9 82.4)
	500	79.4	(75.9 82.9)
	1000	77.4	(74.8 80.0)
Schull	5	100.0	(100.0 100.0)
	50	58.0	(44.3 71.7)
	300	65.0	(59.6 70.4)
	500	66.2	(62.1 70.3)
	1000	66.6	(63.7 69.5)
Tramore	5	40.0	(-2.9 82.9)
	50	72.0	(59.6 84.4)
	300	72.0	(66.9 77.1)
	500	71.2	(67.2 75.2)
	1000	70.0	(67.2 72.8)

Table 7.3: Percent of runs leading to an outbreak and confidence intervals.

From Table 7.3 it can be seen that the size of the confidence interval decreases as the size of the sample increases. This is not unexpected or surprising as by definition a confidence interval will decrease as the sample size increases. We can also note that the percent of runs that lead to an outbreak appears to become more stable as the number of runs increases especially when looking at Schull and Tramore. Again we think this is not completely surprising as the more runs we do the more data we have and the less influence a few outlier runs will have on the

results.

The decrease in confidence interval size and the stabilization of the results with additional runs can be further seen when plotting the size of the confidence interval versus the number of runs and the percent of runs that lead to an outbreak versus the number of runs. For each of the three towns the plots can be seen in Figures 7.1 and 7.2.

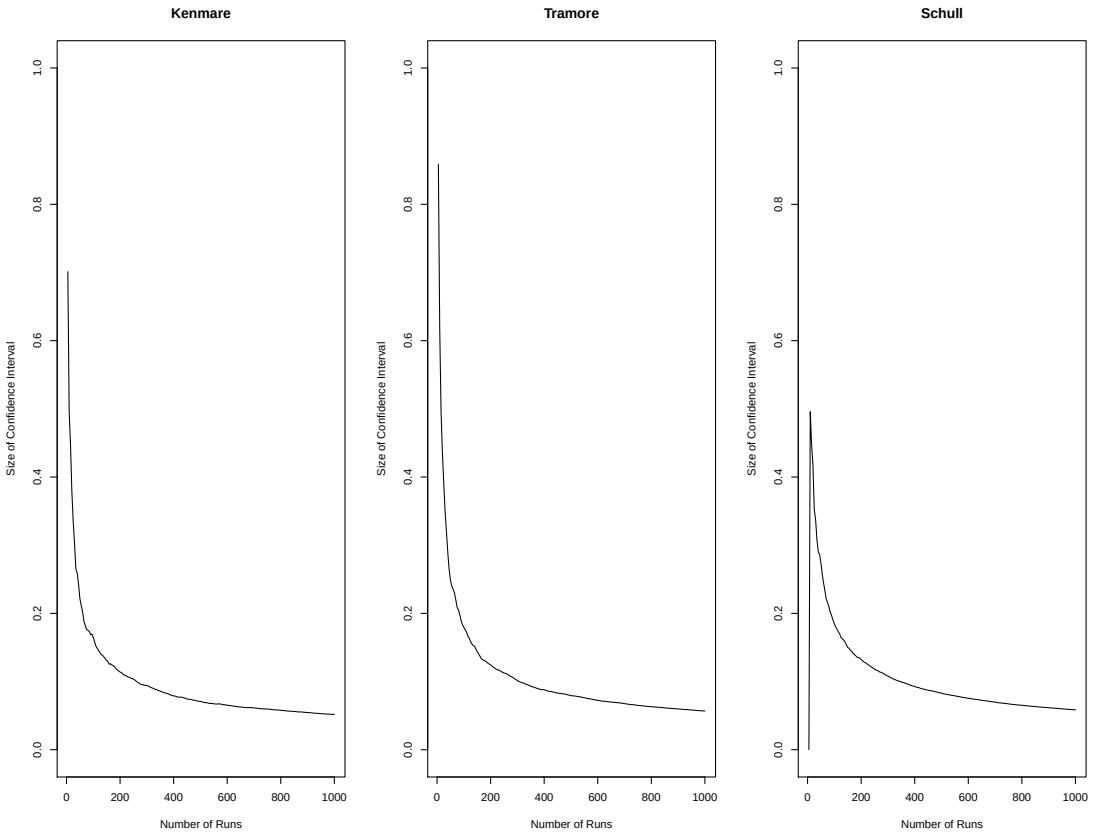


Figure 7.1: Size of the Confidence Interval by Number of Runs for Kenmare, Tramore and Schull

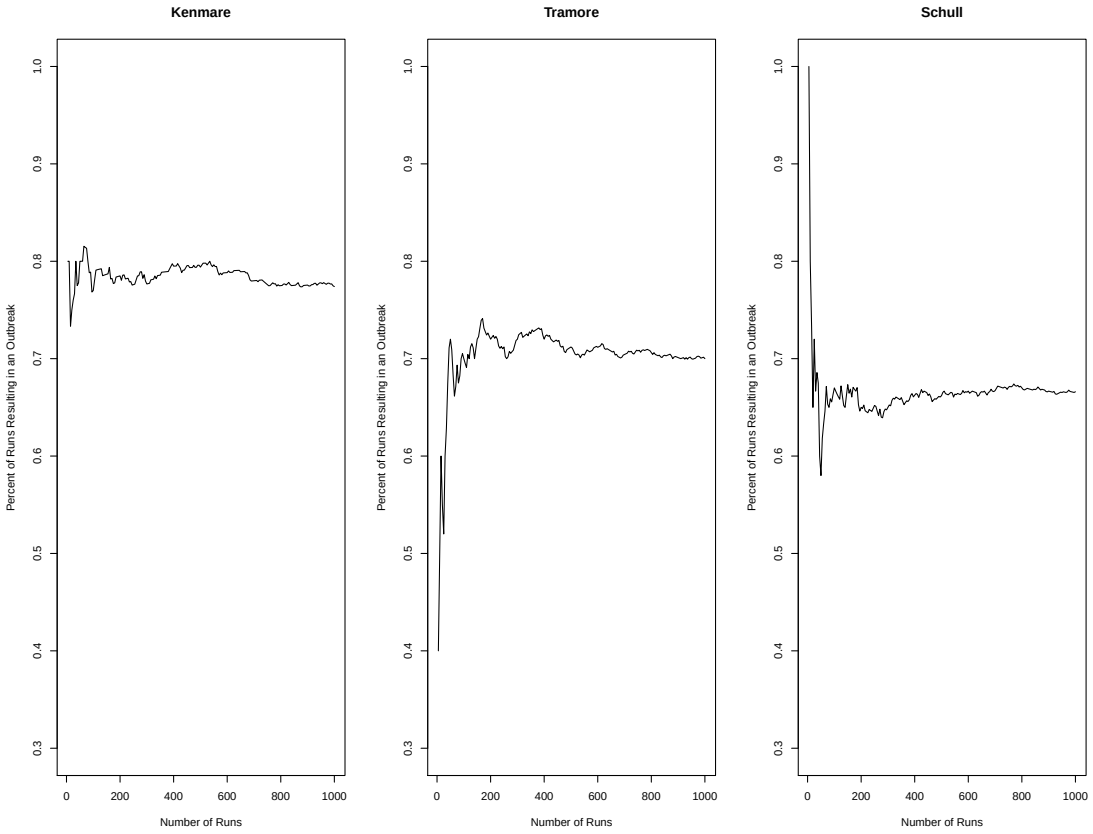


Figure 7.2: Percent of Runs that Results in an Outbreak by the Number of Runs.

7.1.2 Conclusion

As expected the length of the confidence interval decreases as the sample size increases. Looking at the size of the confidence interval as well as the stability of the statistic, we feel that 300 runs is a suitable number of runs for our model. For most towns that we have studied after 300 runs the percent of runs leading to an outbreak becomes steady with minimal variation. In addition, the length of the confidence interval for the towns with 300 runs is approximately 10%. We feel that 10% is an acceptable size for the confidence intervals and upon comparing the results with confidence intervals across 33 towns we still see some distinct differences

in results. Using 1000 runs instead of 300 we cut the size of the confidence interval down to 5%, however, we feel that decreasing the size of the confidence intervals by 5% with an extra 700 runs is not worth the time and computing power that is required by those 700 runs. While the size of the confidence interval is decreasing with additional runs, we are not seeing a large change in the statistic after 300 runs.

As the confidence intervals and parameter stability follow roughly the same pattern for each of the towns we feel that we can make the assumption that using 300 runs is acceptable across all town models. While it might seem like a better option to check the stability and confidence interval size for each additional town that is modelled this will take up unnecessary time and computing power as it is most likely that the test will involve doing additional runs. In addition, while we feel that 300 runs is the correct choice based on our results, the choice is based on opinion and not strict criteria. It is possible that others looking at the same results may chose a different value for the number of runs they feel is necessary to account for the stochasticity in the model. However, not withstanding the fact that interpretations of the data may differ, we believe that the method we have used here, based on confidence intervals is a useful approach to make informed decisions on the number of runs to use for an agent-based model in a given domain.

7.2 Model Validation

Determining a method to evaluate and validate an agent-based model can be difficult as discussed in Section 2.3.5. There is no set methodology for agent based models and often there is no data to compare the results of the model to. Based on the methods used in the literature discussed in Section 2.3.5 we have come up with a basic methodology for validating an agent-based model.

We first consider cross validation: using the results of another previously validated model as a baseline for the agent-based model results. In Section 5.3.1 we compare the results of our basic town agent-based model to the results of a SEIR differential equation based model. SEIR equation based models have been the standard of infectious disease modelling and have been shown to capture the macro level dynamics of an outbreak. Thus if a simple version of our agent-based model is able to capture the same results we can conclude that the basic dynamics of our model are working correctly and that we can then add more complexity to the model as we will in Chapter 8. Once we have validated our simple agent-based model using an equation based model, the agent-based model can become our baseline model for more complex models. This method of validating a model by cross-validating against a simpler model can be generalised beyond cross-validating between an equation based and agent based model. For example, when we scale the model up to the county level in Chapter 9 we will use the agent-based model from Chapter 5 as a baseline to cross validate the town level dynamics of the county model and we will use the county model to cross validate the hybrid model

from Chapter 10.

After cross validation is completed it may be necessary to consider a sensitivity analysis to test the adequacy of the model. A sensitivity analysis allows us to determine if the model is behaving as expected when the values of given parameters change. To do this it is important to determine what parameters should be investigated and what range of those parameters should be used in the analysis. In Section 5.3.2 we do a sensitivity analysis on our town agent-based model. A number of parameters such as the infectivity of the disease, the vaccination rates of the agents and the chance that a sick agent will stay home when sick were explored. The parameters were selected as they were parameters that we had a strong expectation with regards to how the model results should change as the parameters changed. The ranges of parameters chosen were selected so that the ranges included extremes (no agents immune versus all agents vaccinated) and other values that were in the range of possible values. The requirement for sensitivity analysis of a model does not always hold, for example when the model is scaled up to the county level in Chapter 9 there are no new major parameter additions to the model and as it was cross validated using the town model we do not feel it is necessary to perform an additional sensitivity analysis. The hybrid model in Chapter 10 does, however, have a new parameter the threshold switch. As we compare the performance and the results of the hybrid model to the county level model we also look at how the hybrid model performance and results change as the threshold changes showing the sensitivity of the model to the threshold.

The last potential step in evaluation process is comparing the results of the model to existing data. This is most likely the step that will occur least often as it is not always possible to find data that corresponds to what we are modelling. For the model in Chapter 5 we are able to compare the results from a measles outbreak in the town of Schull to a real outbreak in Schull in 2012. However, for later versions of the model we do not have real outbreak data that corresponds to the county that we are modelling.

7.3 Model Testing

The final step in the model evaluation and testing process should be testing the model. A model is only useful if it can help us learn something that we did not already know about the system. Thus it is important as a final step of the modelling process to determine if the model is able to do that. A test needs to be implemented such as the ability of the model in Chapter 5 to simulate the spread of measles in a collection of different towns. In Chapter 9 we will test the role of the centrality of a town in the spread of an outbreak and in Chapter 11 we will implement a school closure intervention strategy to attempt to mitigate an outbreak. These experiments allow us to conclude that not only is our model working as we expect it should but that it is a useful model that can help us to learn about the system and in planning for future outbreaks.

7.4 Conclusion

A methodology for model validation and testing is important when creating a model to show that the results can be used for accurate predictions. In order to be effective a model should be based on appropriate data and should be validated - both of which can prove to be a challenge. Data accessibility is a major obstacle when creating an agent-based model. If agent-based models are to be routinely used as policy tools a consistent validation method should be determined. Without such a method it may be difficult to distinguish a model that will provide accurate results for a given population from a model that will not. However, there is no set methodology in the literature. In this chapter we aimed to create methodology that can be applied to agent-based models for the spread of infectious diseases based on the work we did to evaluate our model in Chapter 5. Abstracting away from the specifics of the models in this thesis, the methodology can be summarized as follows:

1. Use confidence intervals in order to determine the number of runs necessary for the agent based model so as to achieve consistent results.

2. Validation of the model can be done using the following three approaches, and we would recommend using all the approaches that are applicable to a model:

   (a) Cross-validation is useful to validate the basic dynamics of the model when a simpler and well understood model is available.

(b) Sensitivity analysis is useful when there are model parameters for which strong expectations exist with regard to how changing the values of these parameters should affect the results of the model.

(c) Compare the results of the model against existing data where possible.

3. Test that the model is useful by demonstrating that it can be used to test hypotheses.

As the available data and previous existing models can vary depending on what is being modelled, what we propose are only guidelines and can be adjusted to better suit a given model. We feel, however, that our methodology is a useful starting point for an agent-based modeller to help determine the appropriate process for model evaluation. In the chapters going forward, we use the methods outlined here of cross validating against a simpler model, sensitivity analysis and model testing to show that our model is producing accurate and meaningful results.

Chapter 8

Adding Complexity to the Agent-Based Model

One main advantage of an agent-based model is the ability to easily add complexity into the model and make some or all components of the model more realistic. Although the model presented in Chapter 5 was created using real data to simulate a real society, there are additional characteristics of the society that we did not model but could have an influence on an outbreak. Many studies have found that demographic and socioeconomic factors can have an influence on individual health (Mackenbach et al., 2007; Doherty et al., 2014; Jessop et al., 2010; Endrich et al., 2009). Across Europe, there is a trend of higher morbidity and mortality among those with lower levels of income and or education, both factors that contribute to an individuals socioeconomic status (Mackenbach et al., 2007). Although there are many ways in which socioeconomic factors can affect health, in relation to

the spread of infectious diseases one major factor is vaccination rates. A number of studies have shown that various demographic factors can have an effect on vaccination status. Doherty et al. (2014) studied the inequality in childhood vaccination status in Ireland and found that the majority of the vaccination inequality can be explained by socioeconomic variables. They found that socioeconomic status, household structure and equalized income explained most of the inequality in vaccination status. Jessop et al. (2010) studied predictors for the uptake of the first MMR vaccine and determined that a higher level of MMR vaccination was found among children with working mothers, and higher income families. Lower levels of vaccination were found among children with degree level educated mothers, unmarried or lone parents and smoking mothers. A study by Endrich et al. (2009) found that in Ireland there is a lower chance of having been vaccinated for the flu if an individual has more than a primary level education and has a higher income. As many outbreaks of infectious diseases occur today due to low vaccination rates, it is essential to accurately capture this variation in vaccination levels within a population to accurately capture a disease spread. A cluster of individuals with a lower vaccination rate may lead to an outbreak even in a highly vaccinated population. Ireland's Central Statistics Office (CSO) provides detailed demographic statistics on the Irish population at multiple geographic levels, the smallest of which is the small area level that contains information for between 50 to 200 dwellings (CSO, 2014b). While the scale of the CSO small area data might provide information on small areas that have a higher proportion of individuals

with one characteristic over another - for example more single person households versus families with children - there may still be clustering of those individuals or households with those characteristics within the small areas that the data does not capture.

If that clustering is not captured it is possible that the results of the model will not be accurate. More severe outbreaks of non-vaccinated individuals may be missed. On the other side less severe outbreaks due to clusters of vaccinated individuals leading to herd immunity in smaller communities might also be missed. This could create problems when using the results of such a model, for example a vaccination policy based off of model results that included socioeconomic clusters might focus on vaccinating individuals in certain socioeconomic classes while a policy based off a model without clusters might not be as effective.

While synthetic populations that can be used to set up agent-based models with a population of agents that represents the actual population exist for some countries such as the USA (Wheaton et al., 2009), there is no such synthetic population for Ireland. Without access to such a population we need to create our agent population that accurately represents the Irish population. Although the CSO provides rich data on the characteristics of the Irish population, there are some characteristics that are not available such as the level of socioeconomic clustering with in neighbourhoods. In order to account for the possibility of socioeconomic clustering within an epidemiological agent-based model we propose using a segregation model focusing as a burn-in step in the setup of our agent-based

model. The segregation burn-in will allow households in the model to move their home location with in a given area in an attempt to find neighbours with similar socioeconomic status. The burn-in model should help include a factor, socioeconomic clustering, that we feel could be important in an infectious disease model but we do not have the appropriate data to simulate it initially. We believe that this allows better simulation of towns and cities than is possible using summary data at the small area level alone. We apply the segregation model step to an agent-based model created to simulate the spread of infectious diseases in Irish towns. Segregation models are a specific form of the more general social interaction model developed by James Minoru Sakoda (Hegselmann, 2017). Sakoda's model features a checkerboard landscape where two different groups move across the landscape. Agents movements are determined by positive, negative or neutral attitudes towards the other group. Although Sakoda's model includes segregation, segregation models were made famous by a set of models developed by Thomas Schelling in the 1960s and 1970s (Hegselmann, 2017).

As we are only assuming the model does not capture socioeconomic clusters, before adding additional steps into the model, we first must establish that there is a need for using a segregation model. Section 8.1 provides background on segregation models. Section 8.2 discusses the process of using real data on house prices, as a proxy for socioeconomic status, to determine the level of clustering that already exists in Irish towns. This is done by using the dissimilarity index, a measure commonly used to determine a numerical value for a level of segregation in a

given region. The dissimilarity index is calculated for two Irish towns, Schull and Tramore. These towns are selected as examples for the model as they are towns previously used in the infectious disease model. In Section 8.3 the distribution of agents by socioeconomic status in our initial model setup is used to calculate a second dissimilarity index for each town. The simulated dissimilarity index is compared to the real dissimilarity index (calculated in Section 8.2) and is used to determine that an additional step in the model is needed to account for real world clustering by socioeconomic status that we do not capture in the model. Finally in Section 8.4, the segregation step of the model is implemented and again using dissimilarity indices we show that the agent-based model is better calibrated towards socioeconomic clustering after the segregation model has been run. Section 8.3 also discusses the effects that clustering by socioeconomic status has on the outbreaks in each town by comparing two sets of model results one where the initial model setup includes the burn-in segregation model and one without the burn-in segregation model. The experiments done on socioeconomic clustering were previously published in Hunter et al. (2018c).

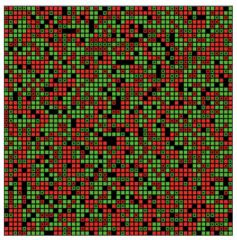
8.1 Segregation Models

Although Schelling presented a suite of models the one that is most famous (the model is often referred to as the Schelling model) is a two-dimensional spatial segregation model. Schelling's model environment is broken up into grid cells in a checkerboard pattern. Two groups of agents are scattered randomly through-

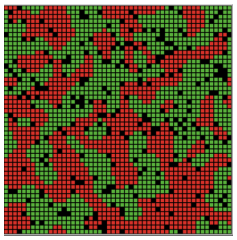
out the checkerboard, each in its own grid cell. The two groups are primarily interpreted as individuals belonging to two ethnic groups, however it is possible to consider other groups. Some grid cells are left empty to allow for movement. Agents have a tolerance for the proportion of similar individuals that they want to live near, with the standard case being that agents do not want to be in the minority so they would have tolerance of at least 0.5. Agents look at their neighbours, agents in the adjacent cells, and determine if the proportion of neighbours who are similar to themselves is greater than or less than their tolerance. If the proportion is equal to or greater than their tolerance the agent will not move. If the proportion is less than their tolerance the agent will move to the closest empty grid cell. Variations include changing the tolerance of agents, the proportion of the population in each group, movement rules, and neighbourhood size (Schelling, 1969, 1971).

Schelling's models illustrate the idea that slight individual behaviours or preferences can lead to aggregate results that the individuals did not intend. Schelling's models show how a small individual preference to not be a racial minority in a neighbourhood leads to neighbourhood segregation (Schelling, 1971). Figures 8.1a and 8.1b illustrate the segregation of two groups of agents resulting from a Schelling type model in Netlogo. Figure 8.1a is the starting point of the model with no clustering or segregation and Figure 8.1b is the ending point of the model where distinct clusters of agents have formed.

Schelling's results have proven to be robust with many other researchers recre-



(a) The setup of a Schelling type model in Netlogo. The image is before any movement has taken place and shows the two groups of agents, in red and green, scattered randomly throughout the environment (Wilensky, 1999b).



(b) The results of a Schelling type model in Netlogo. The image is after the model has run and shows distinct clusters of the two groups of agents (Wilensky, 1999b).

Figure 8.1: Setup and Results of a Schelling type model in Netlogo.

ating and expanding on his work. Stoica & Flache (2014) create a model that extends the idea of residential segregation to school segregation with families picking schools based on both distance and the existing racial mix at the school. Muldoon et al. (2012) investigate the effects changing the utility function or the agents' preferences have on the final segregation of the model. They found that even under conditions where agents prefer to be in a small minority, segregation still occurs when agents have partial information about their surroundings. Survey data showing real residential preferences has been used to show that while different groups have different preferences for neighbourhood make-up, the real world

preferences can be used in a Schelling segregation model that produces the results Schelling predicted (Clark, 1991).

However, real world neighbourhoods are more complicated with other factors than race. For example, in looking at the household survey data, Clark (1991) note that white households will not discriminate against a number of equal status black households, showing that people not only take race into account but other factors such as income and education as well. Clark & Fossett (2008) assert that other factor such as multiple ethnic groups, socioeconomic status, and urban and demographic conditions are needed to truly understand and investigate residential patterns. These factors are not included in Schelling's model. Even when only one factor is considered, a constant tolerance level, which is often used in Schelling type models, is unrealistic. For example, Benenson et al. (2009) use surveys to investigate cases of wealthier households in poor neighbourhoods and find that these wealthier households do not discriminate against their less wealthy neighbourhoods. In some cases there are advantages of the poorer neighbourhoods such as lower house prices.

As the Schelling segregation model only considers a world with simplified features, work has been done to expand the Schelling model to include other factors that influence neighbourhood selection. Fossett (2011) includes not only race in his model but socioeconomic status as well, giving agents preferences for housing quality, neighbourhood socioeconomic status and neighbourhood ethnic mix. Benenson et al. (2009) use tolerance to different income levels as an agent vari-

able in their model and allow it to change with the household when modelling the segregation in Israeli cities.

8.2 Assessing the Degree of Clustering by Socioeconomic Status in Irish Towns

To determine if it is necessary to adjust a model for clustering by socioeconomic status we first must determine if there is evidence of the phenomenon in Ireland. We do not have data of the exact locations regarding where individuals with different socioeconomic status live. However, the Property Services Regulatory Authority provides records in the Residential Property Price Register on the price of properties sold by address from 2010 to the present day (PSRA, 2012). The Residential Property Price Register provides information including the date of sale, address, county and price[1].

Over the past few years there has been emerging literature showing the relationship between property value and socioeconomic status (Coffee & Lockwood, 2012). Moudon et al. (2011) found that neighbourhood wealth measures such as property values had the potential to replace area-level socioeconomic status measures. Coffee & Lockwood (2012) found that a relative location factor based on property value can be used as a proxy for socioeconomic status. Their study determined this factor can be used to enhance area level measures and identify groupings within a given area. If we can consider house price as a proxy for so-

[1]https://www.propertypriceregister.ie/website/npsra/pprweb.nsf/page/ppr-home-en

cioeconomic status then it should be possible to determine if clusters of households of the same socioeconomic status exist in Ireland using the data.

8.2.1 Residential Property Price Register data

For the purpose of finding clusters within the Residential Property Price Register we split the data into six subsets by the sale price of the house. The first subset has the houses with prices in the lower 16.67% , the second subset has houses from the second 16.67% and so forth. Table 8.1 shows the breakdown of prices for each subset of the dataset.

Group	Price Range
Housing Group 1	$<$ 73,000
Housing Group 2	$\geq$ 73,000 and $<$ 120,000
Housing Group 3	$\geq$ 120,000 and $<$ 165,000
Housing Group 4	$\geq$ 165,000 and $<$ 225,000
Housing Group 5	$\geq$ 225,000 and $<$ 320,000
Housing Group 6	$\geq$ 320,000

Table 8.1: House Prices from PSRA (2012)

The entire data set was then geocoded using QGIS (QGIS, 2009) so that it could be loaded into Netlogo (Wilensky, 1999a).

8.2.2 Calculating Dissimilarity in Irish Towns

To determine the level of clustering that exists in a given region the dissimilarity index is used. The dissimilarity index is a measure that determines the "evenness" of a population or the differential distribution of social groups within a region which is composed of a set of *areal units*. It is a popular measure to determine the level of segregation in a space. The dissimilarity index is presented by Massey

& Denton (1988) as one of the main measures to determine segregation. It is also used by the US Census Bureau in determining levels of segregation (Iceland et al., 2002). For the calculation of the dissimilarity index of a region the region is broken up into smaller spatial units call areal units. If any group is segregated then that group will be unevenly distributed. The index produces a value between 0 and 1. The closer the value to 0 the more even and less segregated a region is and the closer the value is to 1 the less even and more segregated a region is. The dissimilarity index is calculated as:

$$D = \sum_{i=1}^{n} \frac{t_i |p_i - P|}{2TP(1-P)} \tag{8.1}$$

where n is the number of areal units in the region the index is being calculated for, t_i is the number of households in areal unit i, p_i is the proportion of minority households in areal unit i, T is the total number of households in the region and P is the proportion of minority households across the whole population of the region (Massey & Denton, 1988).

As our model is run on Irish towns we calculated the dissimilarity index for two towns that we have used our model for, Schull a small town in West Cork and Tramore a town in Waterford. Schull has a population of under 1,000 and Tramore has a population of about 10,000. As the Residential Property Price Register data has only 29 houses sold in Schull and 166 sold in Tramore between January 2010 and February 2017, the dissimilarity index was also calculated for Cork and Waterford cities to determine if the low numbers of houses sales in our

target towns affected the dissimilarity index.

To calculate the dissimilarity index for each town, we break the Netlogo environment up into square grids. This is done by selecting a set of patches that will be the center of each square grid in Netlogo. Then the radius function in Netlogo is used to select patches within a radius of 5 and 10 units of each center patch (1 unit equals 1 patch in the Netlogo environment and approximately 111 m^3 in the real environment being simulated). The radius of 5 produces areal units of 10 x 10 patches and the radius of 10 produces areal units made up of 20 x 20 patches. The square grids become our n areal units. As we are exploring the effects of clustering on the model we use two different radii to determine how the size of the areal units affects clustering. Because the house price group that is in the minority might vary between towns, for each town the dissimilarity in that unit is calculated for each house price group. For example, for the house in the first range p_i becomes the proportion of households in that unit that are in the same price range and P is the total proportion of households in the same price range.

Table 8.2 shows the dissimilarity index for the two towns and two cities calculated with a radius of 5 and 10 units. We investigated a number of other radii but only present results for two as we found a similar patterns and results for all radii. It can be seen from the table that the dissimilarity index for Tramore is similar to that for Waterford. Although there are some differences between the dissimilarity indices they are within a range that could be explained by different make ups in the towns. Schull, however, has larger differences in the less than €95,000 range

and the €95,000 to €165,000 range. This is likely to the limited number of houses in the data set in those ranges. The dissimilarity indices for the towns based on house price data will be used as a benchmark to compare the dissimilarity indices for our simulated towns based on socioeconomic status in order to determine if randomly allocating agents to locations within a small area creates an appropriate distribution of agents by socioeconomic status in the towns.

Price Range	Radius	Schull	Tramore	Waterford	Cork
$< 73{,}000$	5	0.250	0.385	0.359	0.311
	10	0.481	0.319	0.316	0.259
$\geq 73{,}000$ and $< 120{,}000$	5	0.754	0.644	0.600	0.535
	10	0.592	0.558	0.419	0.430
$\geq 120{,}000$ and $< 165{,}000$	5	0.846	0.682	0.626	0.499
	10	0.885	0.529	0.464	0.351
$\geq 165{,}000$ and $< 225{,}000$	5	0.659	0.573	0.696	0.532
	10	0.511	0.566	0.514	0.356
$\geq 225{,}000$ and $< 320{,}000$	5	0.584	0.766	0.782	0.581
	10	0.627	0.682	0.647	0.445
$\geq 320{,}000$	5	0.616	0.882	0.808	0.656
	10	0.565	0.937	0.723	0.513

Table 8.2: Dissimilarity index for each of the housing groups from PSRA (2012)

8.3 Modelling Irish Towns

We start with the model presented in Chapter 5. We do not change the main setup of the model except we include a few additions. We assign working agents a social class again based on CSO data. Social class is the CSO variable used to capture socioeconomic status. Agents are assigned to the classes *Professional workers*, *Managerial and technical*, *Non-manual*, *Skilled Manual*, *Semi-skilled*, and

Unskilled. Households are then given the social class of one of the adults in the house, the adult is randomly selected from all the adults in the household. For household social class *Skilled Manual* and *Semi-Skilled* are combined into one group (*Skilled/Semi-skilled*) and *Unskilled* is combined with unemployed individuals to create the *Other* group. The groupings are based off of those used by Doherty et al. (2014) in their analysis on the effects of socioeconomic status on vaccination rates in Ireland. Irish vaccination data is used to determine the percentage of each age group that received vaccinations and this is adjusted based on socioeconomic status using the odds of having children vaccinated from (Doherty et al., 2014). We also included a retired grouping that was not included in the research by Doherty et al. (2014). Unoccupied households are also added into the model. The number of unoccupied households in a small area is taken from the census data and the households are randomly placed in locations in that small area. When the model is setup we compare the dissimilarity index scores measured in the models of the towns Schull and Tramore with those calculated based on the PSRA. We then describe how the inclusion of a segregation model following Shelling changes this. Schull and Tramore are chosen as they are two towns in Ireland where concentrated measles outbreaks occurred. In 2012 there were at least 30 confirmed cases of measles in Schull in one outbreak (Corner et al., 2012) and in 2013 there were approximately 20 confirmed cases of measles in Tramore in one outbreak (O'Connor et al., 2016).

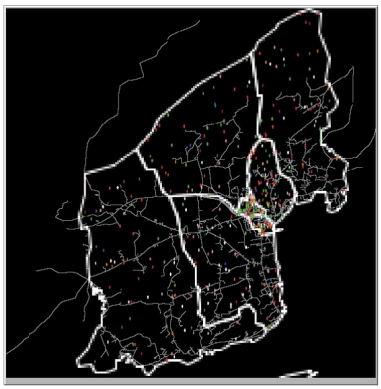
8.3.1 Calculating Dissimilarity In The Basic Town Model

We initiate setup for our model for both Schull and Tramore 300 times (each run will be slightly different due to random initialisation) and calculate the dissimilarity index for each of the household social classes each time the model is initiated. We then find the average of the dissimilarity index across the 100 model setups[2]. The dissimilarity index is calculated twice for each town, once using areal units with a radius of 5 and once using a radius of 10. Similar to how the dissimilarity index was calculated for the real housing price data, for each areal unit the dissimilarity in that unit is calculated for each social class. Table 8.3 shows the average dissimilarity index for both Schull and Tramore for each social class. Figures 8.2a and 8.2b show the initial setup of Schull and Tramore. Agents are color coded by socioeconomic status.

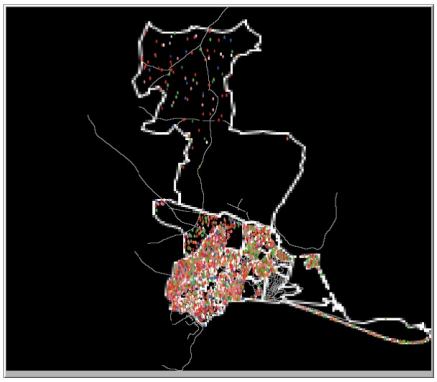
Social Class	Radius	Schull	Tramore
Professional	5	0.378	0.223
	10	0.354	0.202
Managerial and Technical	5	0.346	0.208
	10	0.297	0.151
Non-Manual	5	0.357	0.205
	10	0.297	0.312
Skilled/Semi-skilled	5	0.344	0.206
	10	0.297	0.163
Retired	5	0.346	0.262
	10	0.298	0.263
Other	5	0.359	0.212
	10	0.309	0.168

Table 8.3: Starting dissimilarity index for each of the CSO social class groups

[2]Based on the method in Section 7.1 100 setups results in a confidence interval around the average dissimilarity for each social class of around 0.01 for Schull. We feel this is a small enough confidence interval and will account for the stochasticity in the model.



(a) The initial setup for the town of Schull. The white boundaries are town boarders. Agents are color coded based on their socioeconomic status.



(b) The initial setup for the town of Tramore. The white boundaries are town boarders. Agents are color coded based on their socioeconomic status.

Figure 8.2: The initial setup for the towns Schull and Tramore.

Although house prices and socioeconomic status are not an exact match from the literature we can assume that house prices serve as a proxy for socioeconomic status. If randomly placing houses of different social class within a small area produces realistic neighbourhoods clustered by socioeconomic status we would expect the dissimilarity indices from the real house price data set to be similar to the dissimilarity indices from the model. Comparing the values for the dissimilarity index for each social class from the initial setup of our model to the values for the dissimilarity index for the house price ranges it can be seen that the dissimilarity

indices from the simulation are generally less than those coming from the real data. As the house prices are not similar we can conclude that even when using small area data an agent-based model that uses a random distribution of agents within these areas does not accurately portray neighbourhood socioeconomic status and it may be necessary to adjust the model to account for this. The only exception is the housing price range less than €95,000 in Schull has a dissimilarity index of 0.167 when a radius of 5 is used compared to values of about 0.35 for all the social classes from the simulation. The low dissimilarity index from Schull is likely due to small data samples.

8.3.2 Using a Segregation Model to Better Model Dissimilarity

In order to account for clustering in neighbourhoods by socioeconomic status that we see in our house price data set, but not from the initial setup of the ABM, we add an additional burn-in step into our model setup process. Once the model is populated with agents and all of the agents are assigned the appropriate characteristics we give households the opportunity to move using a Schelling type segregation model. However, unlike in Schelling's models where race is used as the segregating factor we use social class. Each household in the model has a social class assigned to it and households will seek to surround themselves with households of the same social class. To allow this households are given the opportunity to move to more attractive unoccupied houses during the burn-in process.

The model is run on discrete time steps. Each household is considered at each time step. For each household, if the proportion of neighbouring households with the same social class is below a pre-set tolerance level then the household will move to a location with more neighbouring households matching its social class. If the proportion of neighbouring households with matching social class is above the threshold the household does not move. Households moving to new locations is enabled by the inclusion of unoccupied houses in our model setup process. When a household moves they move to the unoccupied house in their current small area with the highest proportion of neighbours matching their own social class. If a better location than their current one is not available households do not move. The burn-in process stops after a time step of no households moving.

The burn-in process involves two parameters: neighbourhood size and a tolerance level for households of different social class. For neighbourhood size we use the same radii used for calculating dissimilarity (see Section 8.2.2): 5 units and 10 units. The next section describes experimental results that are used for setting the tolerance level.

Once the segregation model has stopped running, agents who are students are assigned to the school that is the closest distance to their house and the disease model can be run.

8.4 Results

This section presents the results from two experiments. First we look at how the dissimilarity index for the two towns changes as we adjust the radius and tolerance of agents in the model. The second experiment determines how the clustering affects the results of the infectious disease agent-based model.

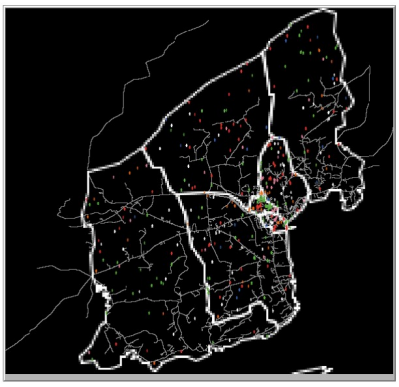
8.4.1 Using Segregation Modelling to Model Socioeconomic Clustering

In this section we report an experiment that tested whether the use of a socioeconomic segregation model improved the calibration of an agent-based model in terms of making the dissimilarity index of the neighbourhoods within the model after the segregation process has been run more similar to real data than the dissimilarity index prior to the segregation model being run. In order to run a socioeconomic segregation model as part of an agent-based model setup process we must set 3 hyper-parameters: the number of iterations the segregation model runs for[3], the tolerance level used in the model, and the radius used in the model. To explore the interactions between the hyper-parameters of the segregation model and the affect of the segregation model on the calibration of the model we: fixed the number of iteration to be 100 and set up a grid search process where tolerance took the values from 0.1 through 0.7 and the radius parameter was set to either 5

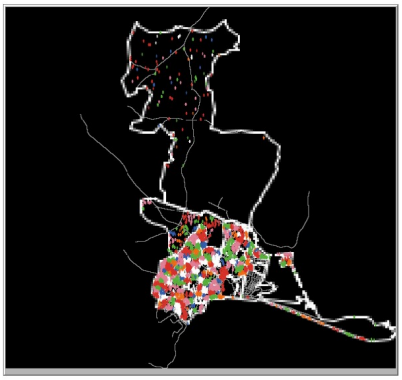
[3]100 runs is selected as the number of model runs necessary based on the method in Section 7.1.

or 10. For each combination of hyper-parameters an agent-based model of Schull and an agent-based model of Tramore was created and a socioeconomic segregation model burn-in process was run in the burn-in process. After each iteration, the dissimilarity index for each social class within each town was calculated and stored and an average was taken across the 100 iterations. Figures 8.3a and 8.3b show the setup for Schull and Tramore after the burn-in model. Clusters of households by color can be seen in both towns, however, as Tramore has a larger population, the clusters are more distinct in Tramore.

Table 8.4 gives the average dissimilarity index for the managerial and technical social class for each of the combinations of tolerance and radii across the 100 iterations. Comparing the final dissimilarity index for each town with the starting dissimilarity index and the real dissimilarity index based on housing price data, it can be seen that the final dissimilarity index is closer to real dissimilarity index. Although we only present the results for one of the social classes in Table 8.4, Figures 8.4 and 8.5 present the results for each of the social classes for Schull and Tramore respectively. The plots in each figure show how the dissimilarity index changes as the tolerances increases. Dashed lines are the starting dissimilarity index while solid lines represent the final dissimilarity after the burn-in process. This can be seen further in Tables 8.5 and 8.6. The two tables show the starting and ending dissimilarity index for each social class in Schull and Tramore respectively. The model used to find the dissimilarity index for Tables 8.5 and 8.6 had a radius of 5 and a tolerance of 0.5.



(a) The setup for the town of Schull after the burn-in has been implemented. The white boundaries are town boarders. Agents are color coded based on their socioeconomic status.



(b) The setup for the town of Tramore after the burn-in has been implemented. The white boundaries are town boarders. Agents are color coded based on their socioeconomic status.

Figure 8.3: The setup for the towns of Schull and Tramore after the burn-in has been implemented.

Within a town the changes in the dissimilarity index based on tolerance and radius follow a similar pattern for each social class. However, between Schull and Tramore the differences are greater. This is likely due to the difference in size between the two towns. Tramore has ten times the population of Schull and thus more households and a greater distribution of social classes. For both towns using a radius of 5 results in higher levels of dissimilarity than a radius of 10. However, for Schull the difference between the starting dissimilarity and

the ending dissimilarity is much greater for radius 10. As the tolerance increases for the Schull model the differences between the dissimilarity for radius 5 versus radius 10 decreases. For most social classes in the Tramore model, as can be seen in Figure 8.5, the difference between results from radius 5 and radius 10 tend to be smallest using a tolerance of 0.2 and 0.3 and then as the tolerance gets greater the difference increases.

Social Class	Radius	Tolerance	Schull	Tramore
Managerial and Technical	5	0.1	0.350	0.211
		0.2	0.355	0.257
		0.3	0.365	0.323
		0.4	0.373	0.342
		0.5	0.377	0.346
		0.6	0.387	0.355
		0.7	0.388	0.355
	10	0.1	0.305	0.156
		0.2	0.322	0.211
		0.3	0.346	0.301
		0.4	0.360	0.330
		0.5	0.357	0.331
		0.6	0.368	0.332
		0.7	0.373	0.335

Table 8.4: Dissimilarity index for the Managerial and Technical CSO social class after segregation model burn-in process using different hyper-parameters.

Social Class	Starting	Ending
Schull		
Professional	0.286	0.391
Managerial and Technical	0.347	0.377
Non-Manual	0.358	0.386
Skilled/Semi-skilled	0.345	0.378
Retired	0.348	0.373
Other	0.360	0.390

Table 8.5: The starting and ending dissimilarity index for the all CSO social classes for Schull. The model used had a tolerance of 0.5 and a radius of 5.

Social Class	Starting	Ending
Tramore		
Professional	0.223	0.334
Managerial and Technical	0.208	0.346
Non-Manual	0.202	0.343
Skilled/Semi-skilled	0.209	0.339
Retired	0.262	0.347
Other	0.210	0.346

Table 8.6: The starting and ending dissimilarity index for the all CSO social classes for Tramore. The model used had a tolerance of 0.5 and a radius of 5.

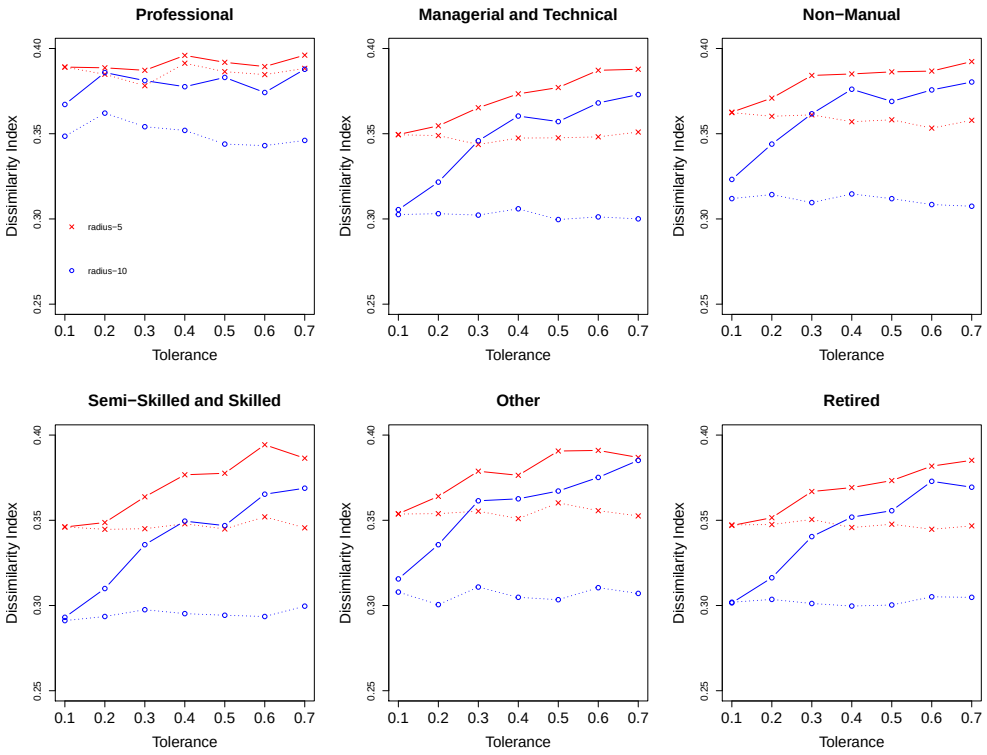


Figure 8.4: Schull results for different levels of tolerance. The dashed lines are average initial dissimilarity index while the solid lines are average ending dissimilarity index.

Although in both towns the simulation still provides a lower value for the dissimilarity index compared to the house pricing data, we feel that the increase in value for the dissimilarity index especially for Tramore moves towards a more realistic artificial society for our simulation. The model shows it is possible to use a segregation type model as a step in the setup of an agent-based model to make

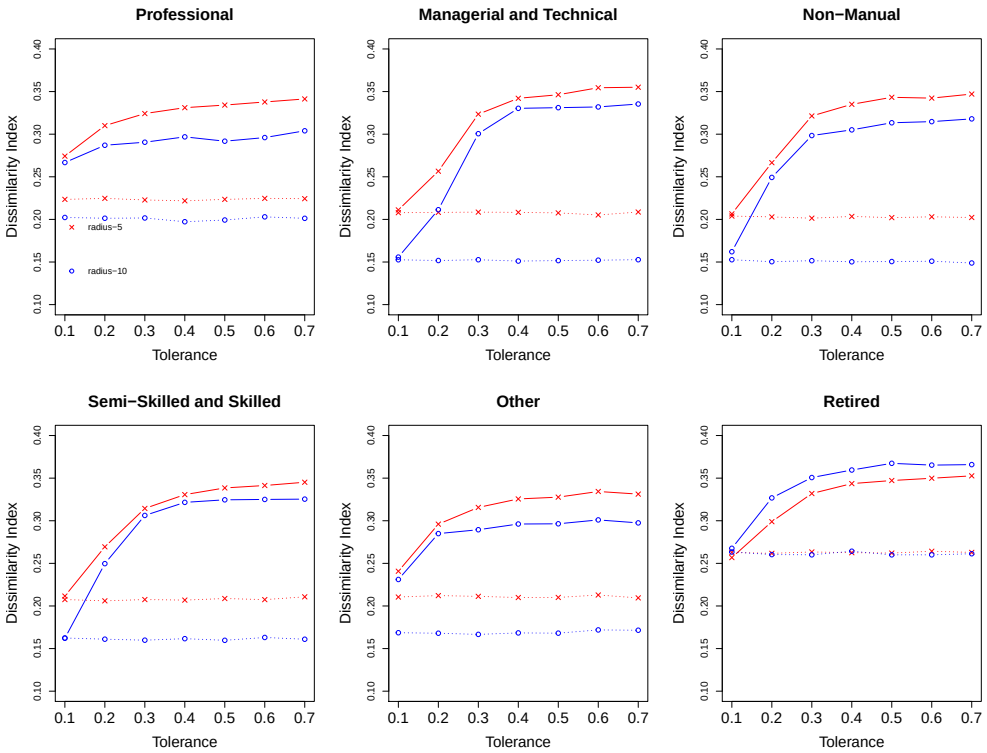


Figure 8.5: Tramore results for different levels of tolerance. The dashed lines are average initial dissimilarity index while the solid lines are average ending dissimilarity index.

the model more realistic. One of the factors causing the differences in clustering could have to do with the extra retired social group that we added. This was done since social class was based on employment status in our model. However, in reality the socioeconomic status of retired individuals might be based on their socioeconomic status before retirement.

8.4.2 Assessing the Impact of Socioeconomic Clustering on Outbreak Modelling

Although it is useful to show that the segregation model does in fact lead to a model that includes socioeconomic clustering, it is important to determine if it has an effect on our infectious disease model. If clustering agents by socioeconomic status has no affect on the results of the final epidemiological model it will not be a useful addition. In order to determine what influence the clustering has on the end results of the model, the infectious disease model was run 300 times with the socioeconomic segregation model as the final steps in setup. For the socioeconomic segregation model a radius of 5 is used with a tolerance of 0.5. The model is run for both Schull and Tramore and results are compared to model runs without clustering. Tables 8.7 and 8.8 show the summary of the results across the 300 runs for Schull and Tramore respectively.

	Schull No Clusters	Schull Clusters
Minimum	1	1
1st Quartile	1	2
Median	15	17
Mean	18.34	20.91
3rd Quartile	20.00	33.25
Maximum	84.00	82.00

Table 8.7: Distribution of total infected agents across the 300 model runs for Schull

Comparing the results for Schull it can be seen that the distribution of total number of infected cases is similar for both the runs with and without clustering. In fact the clustering leads to a slightly larger median and mean compared to results with no clusters. Tramore, however, shows a greater difference between the

	Tramore No Clusters	Tramore Clusters
Minimum	1	1
1st Quartile	2	2
Median	87.0	107.5
Mean	93.7	105.2
3rd Quartile	164.0	177.0
Maximum	459.0	497.0

Table 8.8: Distribution of total infected agents across the 300 model runs for Tramore

runs with and without clusters. With the clusters the magnitude of the outbreaks are greater. This can be further seen looking at the histograms in Figures 8.6 and 8.7. The histograms show the distribution of the number of agents infected in the outbreak. For both distributions there is a higher percentage of runs with a larger number of agents infected when there is clustering compared to the distribution when clustering is not included.

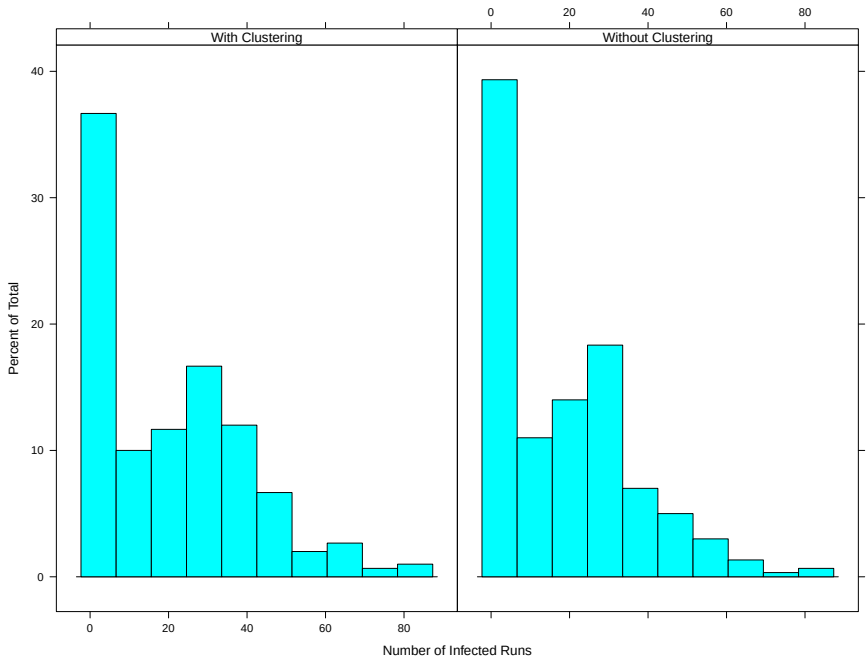


Figure 8.6: Distribution of total infected agents across the 300 model runs for Schull.

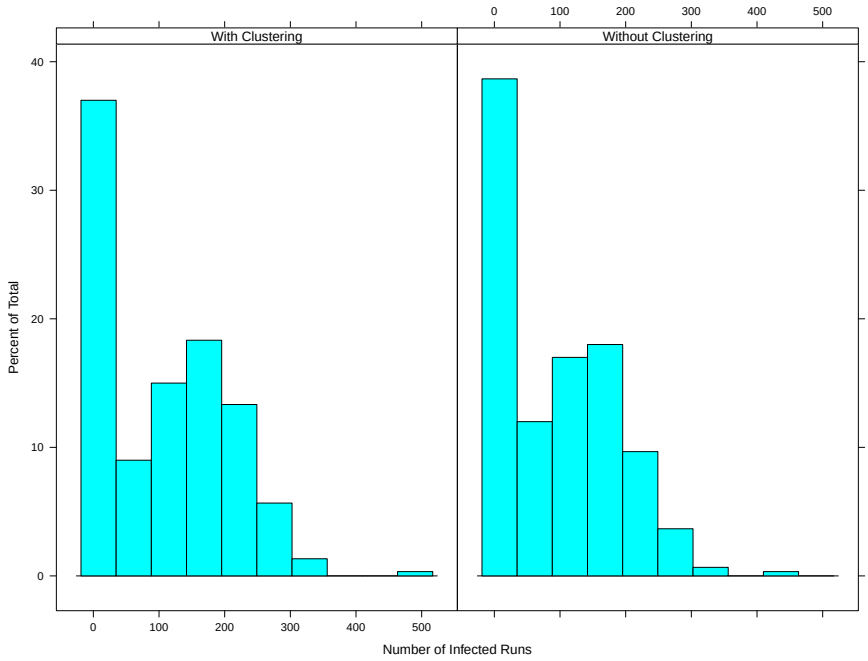


Figure 8.7: Distribution of total infected agents across the 300 model runs for Tramore.

The results are not completely unexpected as Schull is a small low density town and the segregation model does not make a large difference in the level of socioeconomic clustering. For example, from Tables 8.3 and 8.4, the dissimilarity index for the managerial and technical social class with a radius of 5 is 0.346 before the segregation model and 0.377 after the segregation model. This is compared to the Tramore where the dissimilarity index before the segregation model is 0.208 and after is 0.346. Thus it makes sense that the segregation burn-in model does not have much of an influence on the outcome of the infectious disease model for Schull but does have an effect on Tramore. Intuitively the greater magnitude of the Tramore outbreaks also makes sense as in the clustered Tramore model, agents with similar vaccination rates are living closer together and thus interacting more

leading to more infections. Since students choose the closest schools to their home that should also lead to increased infections if students with the lower vaccination rates due to their socioeconomic status attend the same schools. As Tramore has multiple schools for students to attend while Schull only has one primary and one secondary school this could also be a factor as to why there is less of a difference between the Schull runs. The results show that the socioeconomic segregation model can have an effect on the outcome of the model.

8.5 Conclusion

In this chapter we were able to successfully use a segregation model to create a more realistic distribution of socioeconomic status within small areas in order to setup our agent-based model. Through using house price data we determined that clusters by socioeconomic status exist in Ireland and that by randomly placing households within a small area we did not capture the correct level of clustering. Not having the appropriate mix of socioeconomic status could have an effect on an infectious disease model as the overall neighbourhood health and vaccination level of a neighbourhood has an effect on individual health.

Not only have we shown that we can use a segregation model as a step in the setup of our agent-based model for infectious diseases but we have determined that clusters of individuals by socioeconomic status can have an effect on the outcome of the overall model. While Schull does not show this effect, we believe this is due to characteristics of the town. A small town with less than 1,000 residents

may not have enough of a sample of individuals in different socioeconomic groups for clustering to make a difference. In addition, we feel that the school settings should have a large impact on the results of the model. If a school is located in an area that has lower vaccination rates an infectious disease might spread through the school quicker than if there was an equal distribution of vaccination rates with in the school. As there is only one primary and one secondary school in Schull the distribution of vaccination rates will not change regardless of the level of clustering. The increased magnitude of outbreaks in the Tramore runs lead to the conclusion that socioeconomic clustering can result in a different outbreak pattern. Tramore has a much greater population than Schull, with just under 10,000 residents allowing more distinct cluster of agents by socioeconomic status to form and allowing the spatial distribution of agents to contribute to the model results. Thus leading to the conclusion that socioeconomic distribution should be considered as a factor in an agent-based model for infectious diseases.

It is impossible to make a direct comparison between the house price data set, used to determine if clustering exists, and the simulated socioeconmic data set due to data availability. The data desired to create a model is not always available. For example, while we can find house prices for sold houses we do not know the characteristics of the individuals living in those houses. In addition, we have information on individual characteristics but do not have their house prices or exact location. We feel that to move forward in the field of agent-based models it is necessary to take limited data and use assumptions, such as the comparison

between house prices and socioeconomic status, to fill in the gaps of available data.

In addition, it should be noted that while the dissimilarity index is a measure commonly used to determine the level of clustering, it has some disadvantages. It is measured using non overlapping areal units and clustering across areal units is not taken into account with this measure. As an initial exploration into clustering within small areas, and because we are looking for clustering within and across the small areas (each small area has an accurate distribution of households by socioeconomic status before the burn-in model) we feel that the measure gives an acceptable comparison of the clustering before and after the burn-in. However, further work can be done using other spatial methods of clustering such as autocorrelation that might prove the results more robust.

Although, the segregation burn-in model is able to provide us with enhanced results it does come at a price. The detailed town model is already computationally intensive for larger towns and the added burn-in step increases the computational power needed and the time it takes to do a single run of the model. Thus the trade off between detail and time must be considered when running the model.

Chapter 9

Scaling up the Agent-Based Model

The models in the previous chapters create an agent-based model capable of simulating the evolution of outbreaks within a population of a town when the town is considered in isolation. Although information on isolated towns does allow us to analyse the importance of a large number of variables; the fact that the model only considers towns in isolation means that it omits important factors such as travel and commuting patterns between the populations of different nearby towns.

A larger model that considers a network of towns in a given region would be better suited to understand the dynamics of an outbreak and the true susceptibility of a given town. In this chapter we scale up the model to the regional level focusing on counties in Ireland and look into the level of detail necessary to increase the size of the population being modelled without significantly altering the results. A

larger scale model will allow us to learn more about how an outbreak will spread between towns which could be an important factor in stopping an outbreak before it spreads. A larger scale model can also help to identify towns that might have a higher susceptibility to an outbreak than others. One factor that we think might play an important role in both the spread of an outbreak and the susceptibility of a given town is the centrality of the town in the network. The chapter is structured as follows: in Sections 9.1 through 9.5 we first describe how we scale up the agent-based model from Chapter 5 to cover a county in Ireland and then report on a experiment designed to validate this scaled up model; then, in Section 9.6 and 9.7 we used the scaled up model to run experiments to determine the importance of centrality of a town on its susceptibility to an outbreak.

We focus on scaling up this model, from modelling towns in isolation, to take into account the interactions between populations in different towns within a region. In the following sections we break our model down into the four main components of an agent-based model taxonomy outlined in Chapter 4 and discuss the data used to create each component along with the assumptions necessary to scale up the model. A more detailed description of the model in this chapter using the ODD protocol (Grimm et al., 2010) can be found in Appendix B.

9.1 Environment

Agent-based models allow for a high level of detail. Each agent can have as many individual characteristics as desired and similarly the environment can be rich with

detail. However, the greater detail the more computational power needed to run the model and the longer it will take to run. Our aim in scaling up the model is to reduce the fidelity of the environment without greatly influencing the results of the model. The idea is that there is a level of detail in the model that influences our results and there is a level of detail beyond which the results are not affected. In the model in Chapter 5, the environment is made up of small areas and agents can move through the small areas[1] and the town. Residential, commercial and community spaces are designated using zoning data and are used to determine the location of agents houses, workplaces and movements within a small area. In the reduced environment fidelity version of the model, we decrease the environmental space of our model instead of having agents live and move through a small area, each small area is represented by a single point in the model. Each environmental point or *patch* in the model that represents a small area has information about that small area that any agents in the small area can access. This information includes the number of primary and secondary schools in the small area, the number of workplaces and the distances between the small area and every other small area in the model. To test the affect of reducing the environmental fidelity of the model on the model's output, we run the model for two towns Schull and Tramore 300 times with the original environment and 300 times with the reduced environmental fidelity. All other parameters are held constant and the results are compared. Table 9.1 shows the results for the two versions of the model for both Schull and

[1]The smallest geographic area that the Irish census is aggregated over. Each small area contains between 50 to 200 dwellings.

Tramore and the 95% confidence intervals for both. From the results we can see that the abstraction of small areas to points did not cause the results between the two versions of the towns to vary significantly, the results for the detailed environment are within the confidence intervals for the reduced environment and vice versa. From this result we argue that we can use the reduced environmental version of the model without significantly altering the performance of the model.

Town	Detailed Environment	Reduced Environment
Schull	70.3 (65.2, 75.5)	71.0 (65.9,76.1)
Tramore	72.0 (66.9,77.1)	65.3 (59.9, 70.7)

Table 9.1: Percent of runs leading to an outbreak for the detailed environment model and the reduced fidelity environment model

9.2 Society

The society of the model is created using Irish Census data from the CSO (CSO, 2014a). The CSO data is at the small area level. As we described above, small areas are geographic census areas that contain between 50 to 200 dwellings. For each small area we create a population that reflects the population statistics of that small area including age, sex, household size and economic status. Irish vaccination data is used to determine the percentage of each age group that have received vaccinations for the infectious disease being modelled the same as the model in Chapter 5. Social networks are included in the model. Agents have a family social network that is made up of any agents living in their household. Agents also have a work or school social network that is made up of other agents in

their workplace or their school and students are given an additional social network which is a class network that is made up of agents who are in their school and of the same age. Social networks help to determine the contacts an agent has in the model.

In the Chapter 5 model, the population and thus the initial conditions vary slightly from run to run. Each run the model recreates the population again using the same probability distributions. This method allows for variation in the synthetic population and does not settle on a specific version of the population when the exact actual population is unknown. Although it might capture variability in the runs due to one particular set-up being more susceptible than others, there are some disadvantages of running the model this way. The first is time, model set-up can often take a large part of the runtime of the model, holding a population constant can allow for a speedier set-up as the model does not have to make decisions each set-up. In addition, it increases the variability in the output making it difficult to attribute the difference in the output from the runs to agents actions versus the variability of the disease itself. An alternative to this is creating the population once and then using the exact same initial population for each run. If we are attempting to show how different interventions such as vaccination rates influence our results holding the population steady allows us to more accurately attribute changes in our results to the interventions considered.

To test the effect of holding our population constant we take two towns in Ireland, Schull and Tramore, and run the model 300 times changing the initial

set-up each time and 300 times holding the set-up steady. All other parameters are held constant. The results are then compared between the sets of runs. We look at the percent of runs that lead to an outbreak and the distribution of agents infected across runs. We do not expect the results to be exactly the same when the population is held steady versus when the population changes as we will not capture all possible distributions of the town. We do, however, feel that the results should be similar with the results for the changing population showing more variability when compared to the runs with the population held constant. Table 9.2 shows the percent of runs leading to an outbreak for each version of Schull and Tramore and the 95% confidence interval for each percent.

Town	Varying Initial Population	Constant Initial Population
Schull	70.3 (65.2, 75.5)	65 (59.6, 70.4)
Tramore	72.0 (66.9,77.1)	69.3 (64.1,74.6)

Table 9.2: Percent of runs leading to an outbreak for a steady versus a varying initial population

From the table we can see that our results do not vary significantly when we change the model from creating a new population every run to keeping the same population. The results are broken down further by looking at the summary statistics for the number of agents infected each run which can be seen in Table 9.3.

From the distributions it is clear that when the population changes with each start there is a larger variation in the results. We see this with a higher standard deviation in both the Schull and Tramore models as well as higher medians, means and 3rd quartiles. This larger variation in results is what is expected as the varying

Town	Schull Varying	Schull Constant	Tramore Varying	Tramore Constant
Min	1.00	1.00	1.00	1.00
1st Quartile	1.00	1.00	1.00	1.00
Median	5.00	3.00	11.00	5.00
Mean	10.22	5.27	27.73	21.39
3rd Quartile	17.00	8.00	47.25	36
Max	63.00	29.00	129.00	217.00
Standard Deviation	11.26	5.29	33.89	29.51

Table 9.3: Summary statistic for the number of agents infected per run for a steady versus a varying initial population

initial conditions are likely capturing some set-ups that are more susceptible to an outbreak than others. However, our results for the percent of runs that lead to an outbreak do not change significantly between the two versions.

9.3 Transportation

As we described above in Section 9.1, the agent-based model described in this paper differs from Chapter 5 by abstracting away from the small areas level of detail, and so agents do not move within a small area (although, importantly for this work, small areas may now be located in different towns) and the only transportation that occurs is between small areas. Within a small area, all agents in the small area are physically in the same location but the agents keep track of their location within the small area. For example, an agent will know if they are home, at work, at school or in the community. Agents move between small areas but do not move around within a small area. However, all agents in the same small area are not in contact with each other at all times. Instead a variety of factors determine if two agents come into contact with each other. First, is the agent's location, an agent at home will not come into contact with another agent who is

at work. Second is an agent's social networks. An agent will have a higher chance of coming into contact with a member of their family network in the community then a member of their class, school or work network with whom, in turn, they have a greater chance of coming into contact with then an agent who is not in any of their networks.

Between small area movement is modified for the scaled up version of the model. In the town model, agents move in a straight line between their current location and their destination. When they are deciding on their destination within the community the agents will choose randomly from the possible community spaces in town and move there. Although this is an acceptable assumption for a smaller town, when the size of the area being considered increases the assumption does not hold as well. Moving from one side of a town to another in an hours time is not unbelievable, however, moving from one side of a county to another in a short period of time is much less likely. To account for this the agents use a gravity model to choose their next location. Gravity models are a type of transportation model that are based off of Newton's gravitation model. A traditional gravity model gives the interactions between two location pairs and determines those interactions based on the characteristics of a location and the distance between locations (Rodrigue et al., 2006). In the model, agents move between home and school or work at certain predetermined times and will return home at predetermined times. On weekends, summers for students and after school or work hours agents movements through the community are determined by the gravity model. The probability of an agent

moving to another small area is proportional to the population density of the small area, an area with more agents is more attractive, and inversely proportional to the distance to the small area from the agents current location, areas that are farther away are less attractive. We feel that this transportation model provides a more accurate model of movement within a larger area than that in the original town model.

To capture accurate commuting patterns when agents are not moving around the community the CSO Place of Work, School or College - Census of Anonymity Records (POWSCAR) data is used (CSO, 2017). This dataset provides information on the commuting patterns of people in Ireland and gives the number of people that commute from one electoral division to another. Electoral divisions are the census geographic area one step above the small areas.

9.4 Disease

The disease model is based off of a compartmental Susceptible, Exposed, Infected, Recovered (SEIR) model. Where agents start in one of four different compartments, they are either susceptible, exposed, infected or recovered and based on their actions they move between the compartments. It remains unchanged from the model in Chapter 5. When an infectious agent comes into contact with a susceptible agent they have a given percent chance of passing the disease to the susceptible agent. If they do pass on the disease, the susceptible agent then moves to the exposed state for a given period of time before moving to the infectious

state. The will remain in the infectious state for a set period of time before recovering. Once recovered they can not be reinfected. The disease dynamics are set to mimic measles.

9.5 Model Evaluation

The extensions and variations of the model in Chapter 5 that we described above have been designed to enable us to scale the model up to simulate a geographical region, at a county scale, that contains multiple towns. As an initial sense-check of this scaled up model we decided to compare the results for simulations of outbreaks within two towns in the scaled up version of the model with the results for simulations of outbreaks in the same two towns when the towns were isolated within the simulation. Our expectation was that if the scaled up simulation was working appropriately then the outcomes of the simulations under these different conditions should be somewhat different but not drastically so. To do this we run the region model for the county of Leitrim in Ireland. Based on the 2016 census the county has a population of approximately 32,000 people over an area of 1,590 km^2. The county is made up of 173 small areas. We choose two towns in Leitrim, made up of multiple small areas, and compare the results from the county model to the town model. The idea is that we want the town model to be somewhat stable but still be influenced by being connected to the county. For each of two separate towns in Leitrim, Manorhamilton and Kinlough, we run the county model 300 times with the outbreak starting in that town and only look at the agents who

are sick within in the town (i.e., in this experiment we do not consider agents from outside of the town who are sick). We compare these results to the results for the town model for Manorhamilton and Kinlough (in which the towns are only modelled in isolation) and an additional town model that allows for the agents to commute out of the town but no one is commuting in to the town and the only agents in the model are those that live in the town.

9.5.1 Results

The measure we look at for each of our runs is the percentage of runs that lead to an outbreak in the town. We use the WHO definition of a measles outbreak, which is two or more linked cases of measles. For our models we consider an outbreak to be any run where the initial case infects at least one other agent. Table 9.4 shows the results for the two towns.

Town	Town Model	Town Model with Commuting	County Model
Manorhamilton	66.3 (61.0, 71.7)	41.0 (35.4, 46.6)	52.7 (47.0, 58.3)
Kinlough	48.0 (42.3,53.7)	32.0 (26.7,27.3)	42.5 (36.1, 47.2)

Table 9.4: Results for the town models and the county model

From the results we can see that the town only model that allows for commuting results in fewer outbreaks than the town only model where agents can not leave the town. This makes sense as if the infected agents are commuting outside of the town, once they are outside the town they do not come into contact with other agents and thus can not spread the disease until they return to the town. In addition, the county model results for both towns are somewhere between the

completely closed town model and the town model with commuting. Again, this makes sense as in the county model the agents are not restricted to staying within their town so there is a smaller chance of an outbreak in the town in the county model because in some cases the infected agent will commute out of the town and take the infection with them. The outbreak percentage is, however, higher than for the town model that allows commuting because there are other agents in the model who can become infected keeping the outbreak going. We take this as a sign that the county model is working as it should be and not drastically changing the results from the town model.

9.6 Experiments

After showing that the town model within the county is relatively stable we are able to run experiments on the county regional model. Studying an outbreak on a network of towns allows us to study how the outbreak propagates through a network and what different factors influence that propagation. In particular we look at centrality, both the centrality based on the number of agents commuting into and out of a town, also known as degree centrality, and the geographic centrality of the town to all other towns in a network, also known as closeness centrality, and how the centrality of a town influences the spread of an outbreak from a town and the spread of an outbreak into the town. As there are a number of types of centrality, we do not restrict our analysis to just one but instead look at multiple types of centrality and how they interact. We do two different types of runs to

look at centrality: in the first type we run the model with the outbreak starting in a randomly selected small area within the county and then we look at where the outbreak spreads and how many outbreaks occur in each individual town; in the second type the outbreak starts in a given town and we again look at where the outbreak spreads. We run each type 300 times to account for the stochasticity in the model. We use the county of Leitrim for the model and consider 16 different towns in Leitrim. The towns are: Ballinamore, Carrigallen, Cloone, Dromahair, Dromod, Drumkeeran, Drumshanbo, Drumsna, Fenagh, Keshcarrigan, Kinlough, Leitrim, Lurganboy, Manorhamilton, Mohill and Tullaghan.

9.6.1 Centrality

Degree centrality is the main type of centrality that is used in the experiment. Degree centrality is defined as the number of links between each point in the network, but degree centrality can be defined in multiple ways: total degree centrality includes all links into and out of a node, in degree centrality only counts the links going into a node and finally out degree centrality which only calculates the links going out of a node. To get a full picture of how degree centrality effects an outbreak we look at all three versions, which allows us to determine if it is the individuals coming into a town, commuting out or a combination of both that influences the spread of an outbreak.

To account for the number of agents coming into and out of a town a weighted degree centrality is used. The weighted degree centrality is calculated using a

product of the number of links and the average weight of the links adjusted by a tuning parameter. Equation 9.1 shows the formula for weighted degree centrality with DC_i being the centrality of town i, k_i the number of links into the town, s_i the number of agents commuting into or out of the town and α is the tuning parameter. The tuning parameter is used to determine the strength of the weight and the importance of individual link strength: when the tuning parameter is less than one the centrality measure favours more links into the town. If the total number of commuters is fixed a town with more links will have a higher centrality than a town with fewer links. When the tuning parameter is greater than one the centrality measure favours fewer links into the town. If the number of commuters is fixed a town with fewer links will have a higher centrality compared to a town with more links (Opsahl et al., 2010). For the purpose of this study we use an α less than one and set it at 0.5.

$$DC_i = k_i * (\frac{s_i}{k_i})^{\alpha} \tag{9.1}$$

If in degree centrality is calculated instead of out degree centrality, the weights are only those agents commuting into a town and the links are only the paths agents commuting into the town take. Similarly for out degree centrality weights are only those agent commuting out of the town and links are the paths those agents take.

As degree centrality is not the only factor influencing an outbreak we also look at closeness centrality. Closeness centrality is a measure for how physically central

a town is within the network of towns and real world distances are used to calculate the closeness centrality. The formula is:

$$CC_i = \frac{V - 1}{\Sigma(distance(v_i, v_j))} \tag{9.2}$$

Closeness centrality can be defined as the ratio of the total number of nodes in the network (V) minus one to the sum of the distances of the node (v_i) to every other node (v_j) (Chakraborty et al., 2014). Here nodes are towns and the distances are calculated from the center of one town to the center of the next.

Table 9.5 shows the normalized centrality for each of the sixteen towns. The closer to zero the centrality is the less connected a town is to the network. Looking at the the centralities for each town we can see that using these measures, Tullaghan is the least central town. Manorhamilton, has the highest total degree centrality and in degree centrality but a lower out degree centrality and closeness centrality. Keshcarrigan is the town with the highest closeness centrality but has middle values for degree centrality. One pattern that can be seen across many of the towns is that a lower in degree centrality is paired with a higher out degree centrality and vice versa. Although it is not the case for all towns, this seems logical as it is more likely that people living in a town where there are a large number of commuters coming in will not have to commute out of the town. Similarly, if there are only a small number of commuters coming into the town then it seems more likely that the people living in the town will commute out of town.

Town	Total Degree	In Degree	Out Degree	Closeness
Ballinamore	0.67	0.45	0.91	0.95
Carrigallen	0.47	0.45	0.23	0.49
Cloone	0.16	0.09	0.27	0.64
Dromahair	0.32	0.17	0.51	0.44
Dromod	0.37	0.21	0.63	0.53
Drumkeeran	0.34	0.29	0.27	0.75
Drumshanbo	0.89	0.76	0.70	0.95
Drumsna	0.27	0.04	0.74	0.81
Fenagh	0.26	0.17	0.38	0.96
Keshcarrigan	0.33	0.26	0.28	1.00
Kinlough	0.18	0.18	0.01	0.07
Leitrim	0.51	0.22	1.00	0.89
Lurganboy	0.19	0.02	0.56	0.38
Manorhamilton	1.00	1.00	0.24	0.40
Mohill	0.90	0.87	0.41	0.71
Tullaghan	0.00	0.00	0.00	0.00

Table 9.5: Normalized centrality by town

9.6.2 Town Similarities

When comparing the results for different towns it is impossible to say which factors of the towns lead to different results. Two towns with different centrality might also have a different population size, town structure etc. and thus the question should arise which factor is actually leading to the difference. In order to attribute most of any difference in results to the difference in centrality we calculate a euclidean distance between each town. This allows us to choose similar towns across a number of measures so that we are able to better credit any difference in results to centrality. Each town is represented by a vector of quantitative characteristics: population size, town area (km^2), population density, number of small areas that make up the town, the number of secondary schools, the number of primary schools, the percent of susceptible agents in the town and the percent of agents who

are students in the town. All categories except for the number of secondary schools and number of primary schools are standardized. The euclidean distance is then calculated between each of the 16 towns so that we can compare results between similar towns. Figure 9.1 presents a distance matrix which visualizes which towns are similar based on these categories. The lighter the square the more similar the towns are and the darker the more dissimilar.

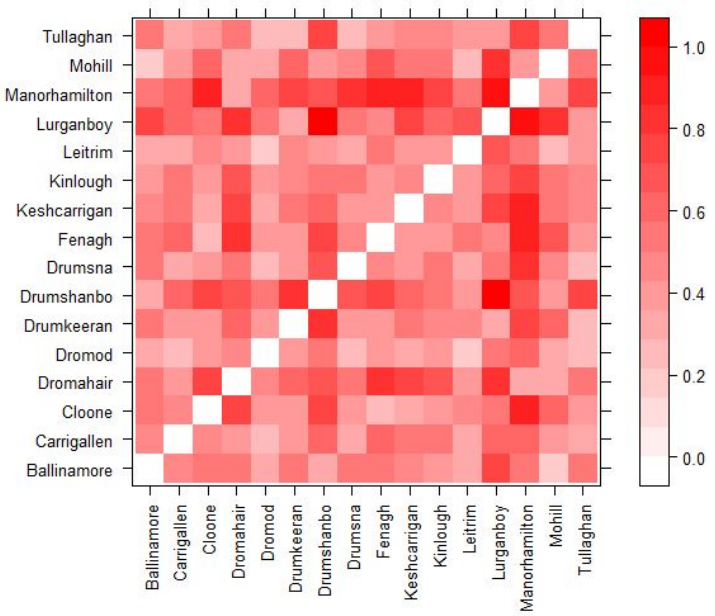


Figure 9.1: Distance Matrix showing the normalized euclidean distance between towns

9.7 Results

For each set of runs we calculate the percent chance of an outbreak (two or more cases of measles) occurring in the town. Table 9.6 shows the percent of runs that lead to an outbreak in each of the 16 towns when the outbreak starts at a

random location in the county, or when it starts in one of eight different towns: Cloone, Dromahair, Fenagh, Kinlough, Leitrim, Manorhamilton, Mohill or Tullaghan. These eight towns were selected to get a range of centralities but were also chosen using the town similarities data discussed in the previous section so that any differences found can be more easily attributed to the differences in centrality.

	Random Start	Cloone	Dromahair	Fenagh	Kinlough	Leitrim	Manorhamilton	Mohill	Tullaghan
Ballinamore	24.0	24.7	20.3	31.3	17.7	29.3	15.7	24.0	19.0
Carrigallen	17.3	25.0	11.3	22.3	13.3	21.7	7.7	16.3	11.0
Cloone	7.7	-	5.3	10.0	6.0	12.3	5.3	15.3	8.7
Dromahair	18.3	19.0	-	23.7	26.7	28.3	31.0	19.3	31.3
Dromod	18.3	27.0	15.3	19.3	13.0	27.7	7.7	22.0	13.0
Drumkeeran	7.3	9.3	14.7	8.3	13.7	8.7	15.0	7.3	10.7
Drumshanbo	18.6	17.7	22.7	23.3	23.0	29.7	18.7	18.0	15.3
Drumsna	14.3	19.7	13.7	12.3	12.0	18.3	8.0	20.3	9.7
Fenagh	11.7	15.3	10.3	-	13.3	15.7	7.3	10.3	8.0
Keshcarrigan	9.3	5.7	10.7	13.7	9.0	16.7	6.0	6.3	8.0
Kinlough	21.0	23.0	24.3	20.3	-	22.3	30.7	17.7	31.3
Leitrim	18.0	19.0	19.3	22.0	16.7	-	16.0	18.7	14.3
Lurganboy	6.7	8.0	14.3	11.3	13.3	18.3	27	7.3	24.3
Manorhamilton	22.7	25.0	34.3	24.7	27.3	28.3	-	18.0	40.7
Mohill	24.3	36.7	19.3	27.3	24.3	28.3	16.7	-	25.0
Tullaghan	15.3	18	18	16.7	20	20	26	16	-

Table 9.6: Percent of runs leading to an outbreak in each of the 16 towns when the outbreak starts in a random location or one of the eight selected towns.

Table 9.6 shows the percent of runs leading to an outbreak in the 16 towns based on where the outbreak began: randomly or in a specific town. Each column represents a different starting location of the outbreak. The results show a range of percents across the different towns. Some towns appear to be more stable than others. For example, Leitrim where the percent chance of an outbreak ranges from 14 to 22 depending on where the outbreak starts, while Dromod ranges from 7.7 to 27.7. One thing to notice with the towns is that the towns that tend to have a more stable percentage change of outbreak despite the starting location are those that tend to have lower centrality scores across all four measures. Similarly, the towns with the lowest average percentage across all starting locations have lower

degree centralities. Table 9.7 shows the Pearson correlations between the different centralities of the towns and the range of percentage of runs that lead to outbreaks for each town based on starting location and the average percentage of runs that lead to an outbreak across all starting locations.

Centrality	Range	Mean
Total Degree	0.51	0.58
In Degree	0.53	0.58
Out Degree	0.02	0.07
Closeness	-0.29	-0.38

Table 9.7: Correlations between centrality and the range and average percentage outbreaks

To analyse correlations we use the following scale: 0 corresponds to no linear relationship, 0 to 0.3 or 0 to -0.3 corresponds to a weak linear relationship, 0.3 to 0.7 or -0.3 to -0.7 corresponds to a moderate relationship and 0.7 to 1.0 or -0.7 to -1.0 corresponds to a strong linear relationship (Ratner, 2009). From the table it can be seen that there is a moderate relationship between total degree centrality and in degree centrality and both the range and mean values. Closeness centrality also has a negative moderate relationship with the mean value a weak relationship with range. From looking at these results it is clear that there is a relationship between centrality and how susceptible a town is to an outbreak spreading to it in a network of other towns. In the next sections we aim to look at the differences in the results by town and how the centrality of the town influences those difference.

9.7.1 Degree Centrality

The results listed in Table 9.7 indicate that total degree centrality is moderately correlated with both the average percent of runs leading to an outbreak across all starting locations and the range of percent of runs leading to an outbreak as well. To look deeper into the relationship between total degree centrality and the results across the towns the correlations between the percent of runs leading to an outbreak in a given town and the total degree centrality of that town are found across all starting locations of the outbreak. These correlations are presented in Table 9.8.

Start of Outbreak	Correlation
Random Start	0.64
Cloone	0.51
Dromahair	0.62
Fenagh	0.68
Kinlough	0.62
Leitrim	0.65
Manorhamilton	-0.12
Mohill	0.34
Tullaghan	0.40

Table 9.8: Correlations between the percent of runs that lead to an outbreak and the total degree centrality of the town by starting location of the outbreak

The results show that for all starting locations except for Manorhamilton there is a moderate correlation between percent chance of an outbreak in another town and the total degree centrality of that town. Manorhamilton is the town with the highest total degree centrality, therefore, we can interpret these results as the higher the total degree centrality of the town the outbreak begins in the less important the total degree centrality of the other towns is to the spread of the

infectious disease. This can be further emphasized when looking at the town with the next lowest correlation, Mohill, which is also the town with the second highest total degree centrality. However, if an outbreak starts in a town that has a lower total degree centrality, the total degree centrality of the other towns has a larger impact on the results.

Because there are multiple parts of total degree centrality, it is made up of both agents commuting into a town and agents commuting out of a town, we break total degree centrality down further into in degree centrality and out degree centrality. This will allow us to understand better what has an influence on an outbreak, agents coming in or going out or if they are both equally important.

In Degree Centrality

Of the different measures of centrality, in degree centrality seems to be the most highly correlated with whether an outbreak will spread to a given town. Logically, this makes sense as an outbreak can only spread to a town if there are agents commuting into the town. Table 9.9 shows the correlations between the percent chance of an outbreak in a given town and the in degree centrality of the town broken down by the starting location of the outbreak.

The correlations show that for all starting locations except for Manorhamilton and Mohill the in degree centrality has a moderate relationship with the percent of runs leading to an outbreak. Both Manorhamilton and Mohill have high in degree centralities while Cloone, Dromahair, Fenagh, Kinlough, Leitrim and Tullaghan

Start of Outbreak	Correlation
Random Start	0.61
Cloone	0.52
Dromahair	0.64
Fenagh	0.62
Kinlough	0.62
Leitrim	0.57
Manorhamilton	-0.08
Mohill	0.22
Tullaghan	0.44

Table 9.9: Correlations between the percent of runs that lead to an outbreak in a town and the in degree centrality of the town by starting location of the outbreak

have lower in degree centralities. This is similar to the results seen in the total degree centrality and thus can be interpreted in a similar way: when an outbreak starts in a town with high in degree centrality the in degree centrality of another town in the network does not have as large of an effect on if the outbreak will spread to that other town compared to when an outbreak starts in a town with lower in degree centrality.

Out Degree Centrality

From the results in Table 9.7 it can be seen that out degree centrality only has a very weak relationship with both the range and mean value of percent of runs that lead to an outbreak. Similarly, if we look at the correlations between the percent of runs that lead to an outbreak and the out degree centrality of the town by starting location of the outbreak in Table 9.10 we can see that if the initial outbreak starts in Fenagh, Leitrim, or Mohill there is a moderate relationship between the out degree centrality of other towns in the network and the percent of runs that lead to an outbreak in that town.

Start of Outbreak	Correlation
Random Start	0.19
Cloone	0.07
Dromahair	0.02
Fenagh	0.29
Kinlough	0.03
Leitrim	0.39
Manorhamilton	-0.16
Mohill	0.41
Tullaghan	-0.20

Table 9.10: Correlations between the percent of runs that lead to an outbreak and the out degree centrality of the town by starting location of the outbreak

Looking further into the results we can see that out degree centrality might have an effect that's not captured in the correlations. For example, when an outbreak starts in a town with high out degree centrality in general the percent of runs that lead to an outbreak in the other towns tend to be higher than when an outbreak starts in a town with low out degree centrality. Leitrim and Mohill are two towns that are similar in most of their attributes with a low euclidean distance between them as can be seen in Figure 9.1 but Leitrim has a high out degree centrality while Mohill has a significantly lower out degree centrality. When the outbreak starts in Leitrim, the average percent of runs that lead to an outbreak in the other towns in the model is 21.7 and it is 15.8 when the outbreak starts in Mohill. Comparing the results town to town shows that for almost every town, they have higher percentage of runs that lead to an outbreak when the outbreak starts in Leitrim versus Mohill. This seems to make sense as a higher out degree centrality would mean that more agents from that town are commuting to other towns resulting in a higher percentage chance of an outbreak spreading.

Compared to the correlations with in degree centrality the correlations for out

degree centrality are markedly different than those with total degree centrality. One possible conclusion from this is that agents commuting into a town are more important than agents commuting out of a town for the spread of an infectious disease through a network. This is likely for a few reason: in the context for measles the most important group in spreading the disease is students and in a smaller county such as Leitrim students are likely to commute into larger towns for school. In addition, agents' behaviours change when they are infected: they become less likely to commute to school or work. Therefore, commuting out of a town will not have as big of an effect on the spread of the disease. Instead it is those agents commuting into a town that will have a greater chance of coming into contact with an infected agent and then spreading the disease to their own home town.

9.7.2 Closeness Centrality

Closeness centrality gives a measure of how close a town is to every other town in the network. Here we focuses on how close in geographic distance the town is to the other towns in the network for closeness centrality as we feel that geographic distance should be an important factor in the spread of a disease: an outbreak should be more likely to spread to a town that is close to a number of other towns than a secluded town in the network as there will be more movement between towns in the former case. Looking at Table 9.7 it can be seen that there is a moderate relationship between the closeness centrality of a town and the average

percent of runs that lead to an outbreak across all starting locations. Looking into this farther, the correlations between the percent of runs that lead to an outbreak and the closeness centrality of that town for each starting location of the outbreak can be seen in Table 9.11.

Start of Outbreak	Correlation
Random Start	-0.11
Cloone	-0.17
Dromahair	-0.28
Fenagh	0.07
Kinlough	-0.27
Leitrim	-0.08
Manorhamilton	-0.63
Mohill	-0.10
Tullaghan	-0.65

Table 9.11: Correlations between the percent of runs that lead to an outbreak and the closeness centrality of the town by starting location of the outbreak

The table shows that two starting locations, Manorhamilton and Tullaghan, lead to a moderate negative relationship between the closeness centrality of the town and if the outbreak will spread there. Two additional towns, Dromahair and Kinlough show a weak to moderate negative relationship. These four towns have the highest closeness centrality, are the farthest away from the majority of towns in the model, of the eight towns that are studied as starting locations of the model and are four of the five highest closeness centralities across the sixteen towns studied. From this we can make an assumption that the less connected a town is when considering closeness centrality, the more important the closeness centrality of the town the outbreak spreads to is. This pattern occurs regardless of the other values for degree centrality, Manorhamilton and Tullaghan both have the highest closeness centrality scores and similar correlations in Table 9.11 but

have different values for total, in, and out degree centrality. Manorhamilton has the highest total and in degree centrality and a lower out degree centrality while Tullaghan has the lowest total, in, and out degree centrality of all the towns. As the towns in the simulation have many other characteristics that could influence the spread of an outbreak, such as population size and age structure, to determine if the relationship between the closeness centrality of the town where the outbreak starts and where the outbreak spreads exists regardless of the other characteristics of the town we look at the euclidean distance between towns from Figure 9.1 along with the closeness centrality and model results. Manorhamilton and Dromahair can be seen as two towns with a closer euclidean distance in Figure 9.1 and the pattern occurs in both towns, while both Tullaghan and Kinlough have farther euclidean distance to Manorhamilton and again the pattern occurs in all three towns. We believe that this shows the other characteristics of the towns do not influence the result that the less connected a town is when considering closeness centrality, the more important the closeness centrality of the town the outbreak spreads to is.

The negative correlation here is interpreted as a higher closeness centrality results in a lower percentage of runs that lead to an outbreak and vice versa. Although an obvious explanation for this is difficult to determine, it would seem that the closer to other towns a given town is it should be more susceptible to an outbreak not less, this result could be down to interactions with other factors and the towns with moderate correlations that are further discussed in the next

section.

9.7.3 Distance and Centrality

Closeness centrality gives a measure of how close a town is to every other town in the network: is the town located near a lot of other towns or is it more secluded. However, a factor that is associated with closeness centrality but not quite captured in it is the distance between the town where the outbreak initially starts and the town where it spreads. For example, if the outbreak starts in Mohill, it is only approximately 12 km to Fenagh but 80 km to Kinlough. It would thus seem that it should be more likely that an outbreak would spread from Mohill to Fenagh than Mohill to Kinlough. Evidence for this can be seen if we look at the results in Table 9.6 and the approximate distances between towns that can be found in Table 9.12.

	Cloone	Dromahair	Fenagh	Kinlough	Leitrim	Manorhamilton	Mohill	Tullaghan
Ballinamore	12	51	5	77	28	56	17	82
Carrigallen	14	66	16	89	34	71	19	93
Cloone	-	55	10	81	25	60	8	88
Dromahair	55	-	50	32	37	14	53	39
Dromod	17	56	21	82	36	61	9	89
Drumkeeran	40	15	35	41	23	20	39	48
Drumshanbo	22	32	17	59	7	37	21	65
Drumsna	18	47	21	73	10	53	11	80
Fenagh	10	50	-	77	18	56	12	84
Keshcarrigan	14	41	8	68	10	46	13	74
Kinlough	81	32	77	-	63	21	78	7
Leitrim	25	37	18	63	-	49	29	70
Lurganboy	62	13	58	20	45	3	61	27
Manorhamilton	60	14	56	21	49	-	58	28
Mohill	8	53	12	78	20	58	-	90
Tullaghan	88	39	84	7	70	28	90	-

Table 9.12: Approximate distances between each town in km

When comparing Tables 9.6 and 9.12 it can be seen that for some cases distance to a town from the starting location of an outbreak has an effect on the results. When an outbreak starts in Tullaghan, a town with high closeness central-

ity and low degree centrality, 41% of runs lead to an outbreak in Manorhamilton. Manorhamilton is a town that is close in distance to Tullaghan, approximately 28 km, and is also a town that has high degree centrality. This is the highest percent of runs that lead to an outbreak for Manorhamilton across all considered starting locations of the outbreak. The next four towns with the highest percent of runs that lead to an outbreak when it begins in Tullaghan are Dromahair (31.3%), Kinlough (31.3%), Mohill (25%), and Lurganboy (24.3%). Dromahair, Kinlough, Lurganboy and Manorhamilton are the four towns closest to Tullaghan that are considered in the model and Mohill has the second highest total and in degree centrality of all of the towns. From this we can consider that when an outbreak starts in a town with low centrality an outbreak is likely to spread to the towns that are both nearby and with those towns with high degree centrality.

Looking across the tables it can be seen that for a number of towns the lowest percent of runs leading to an outbreak come when the outbreak starts in a location that is far away. For example, in Carrigallen the lowest four values for percent of runs leading to an outbreak come when the outbreak starts in Dromahair, Kinlough, Manorhamilton and Tullaghan. These towns are all farther away from Carrigallen then any other town considered in the model. The distance seems to counteract the high in degree centrality for Manorhamilton and the relatively high out degree centrality in Dromahair and protects Carrigallen from an outbreak occurring. A similar phenomenon happens for the town of Dromod where the percent of runs leading to an outbreak when the initial outbreak starts in Kinlough,

Manorhamilton or Tullaghan are significantly lower than when the outbreak starts in other towns. There are also cases of the reverse where towns such as Drumsna and Lurganboy show higher percents when the town where the initial outbreak starts is a town that is close in distance. Lurganboy is very close to Manorhamilton and if the outbreak starts in Manorhamilton there is a 27% chance that the outbreak will spread to Lurganboy. However, if the outbreak starts in Mohill, a town that is similar to Manorhamilton in both euclidean distance and degree centrality an outbreak starts in Lurganboy only 7.3% of the time. The short distance to Manorhamilton clearly plays a key role in the susceptibility of Lurganboy to an outbreak.

Distances might also play a role in the moderate negative correlations found with closeness centrality in Table 9.11. The four towns with a moderate correlation, (Manorhamilton, Dromahair, Kinlough and Tullaghan) all have lower closeness centralities but are close to each other and farther away from the other towns. These moderate negative correlations might be capturing the effects of distance between towns. If a town has a low closeness centrality, the distance to the other towns might be more important to if an outbreak spreads to the other town than the closeness centrality of the other towns.

9.8 Conclusion

Being able to scale up a town based agent-based model to model a network of towns within a region is important in being able to capture and understand the

spread of an infectious disease. No town exists in isolation and movements into and out of a town can greatly influence the susceptibility of a town to an outbreak. As has been shown, assumptions have been made to scale the model from a single town model to a regional model. These assumptions and changes have been shown to not significantly influence the results of the model and allow us to better model a county or region. Although we only model a single county it would be realistic to use the same method to model multiple counties or an entire country. Further work can be done to model multiple counties. Data, and in this case open data, is particularly important because using data allows us to create a realistic agent-based model for the spread of infectious diseases that can be used to help determine the factors that lead to the susceptibility of a given town. When scaling up the model it is important to know what data will make a difference in the results and what data will not. For example, we were able to reduce the environmental fidelity of the model by not including as much GIS data such as zoning data without overly impacting the results. Additional data, however, was used in other areas such as creating a more realistic transportation model.

The scaled up model allowed us to study more than just factors affecting a single town but how the centrality of a town and how connected it was to other towns in the region influences an outbreak in a town. Agent-based models are particularly suited to this task as they capture the interactions of different factors and as we have shown it is not just one type of centrality that impacts if an outbreak spreads to a town but multiple types of connectedness as well as other information about

the network such as the town where the initial outbreak occurred along with the other factors that were identified in Hunter et al. (2018b).

Modelling how agents movements influence the course of an outbreak is important in studying how to react when an outbreak occurs. If an outbreak starts in a given region towns that are more susceptible in that area can be a focus of the response with more resources sent to these towns. In addition, the location of the initial cases can be used to guide responses. For example, if the outbreak starts in a town with high degree centrality closing the town might prevent further spread of an outbreak while if the town has low degree centrality the best course of action might be closing things such as schools or workplaces in nearby towns with high degree centrality as this would stop the outbreak from spreading into the high degree centrality town and then to many other locations from there. Additional simulations could be run including such restrictions to better understand how they could help to reduce the size and severity of an outbreak. In Chapter 11 we will use the lessons learned on how centrality influences an outbreak from this chapter to test intervention strategies involving school closure policies and centrality.

Chapter 10

A Hybrid Model for the Spread of Measles

To reduce the computing power needed to run our agent-based model we propose creating a model that is a hybrid agent-based and equation based model to utilize the advantages of both. From our analysis in Chapters 3, 5 and 6 we saw that an equation based model was able to reproduce a given outbreak and was adaptable to different towns but lacked the stochasticity and emergent results that made an agent-based model attractive. However, it is also these factors that leads to long runtime of agent-based models. The added complexity in an agent-based model for the spread of infectious diseases compared to an SEIR equation based model results in additional computation resources required to run the agent-based simulations. While a typical equation based model and a simple agent-based model can be run on a laptop more complex agent-based models require more computing

power (Bobashev et al., 2007). A hybrid model allows the use of more than one type of model to simulate a system. We hypothesize that there are aspects of our agent-based model that if switched to an equation based model would not significantly reduce the fidelity at the model. In this chapter we develop and test a hybrid model for the spread of infectious diseases.

While the model in Chapter 9 is able to scale up to simulate disease spread in a small county, it takes an extended period of time to run 300 runs of the model and a significant amount of computing power. As we made a set of assumptions in scaling up the model from town to county level that did not significantly alter the results but led to a trade off between fidelity and computing power, this chapter explores further assumptions and trade offs concerning when the model needs to be agent-based versus equation based. The following section discusses hybrid models in more detail. Section 10.2 then presents our hybrid model that simulates an outbreak in a single town. Section 10.3 presents a scaled up version of that model that simulates an outbreak in a county. A more detailed description of the model in this chapter using the ODD protocol (Grimm et al., 2010) can be found in Appendix C.

10.1 Epidemiology Hybrid Models

Hybrid agent-based models are a way to combine the advantages of the "top down" equation based model and the "bottom up" agent-based modelling method while reducing the limitations of both (Marilleau et al., 2018). The hybrid allows for

further scaling and modelling of a larger population while still keeping a heterogeneous population.

Although not abundant in the literature there are some examples of hybrid agent-based models for infectious disease epidemiology. These hybrid models tend to fall into two major categories, a system where some parts are modelled using agents and other parts are modelled using an equation based model and a system that uses both an equation based and an agent-based model and switches between the two (Binder et al., 2012). The model by Bobashev et al. (2007) is an example of the latter. They use a hybrid model to study pandemic influenza. The model is made up of cities in a network with transportation in between the cities. When a city reached a certain number of infected agents it switches to an entirely equation based model. Kasereka et al. (2014) create a model that falls into the former category where agents move between cities based on an agent-based model but the disease model is an equation based model. Similarly Yoneyama et al. (2012) use a equation based disease in their hybrid agent-based model. Hybrid models are also used in infectious disease epidemiology for non-human based diseases. Bradhurst et al. (2013) and Bradhurst et al. (2015) create a model for the spread of foot and mouth disease in livestock. In the model with-in herd disease dynamics is modelled using an equation based model and between herd dynamics is modelled using an agent-based model. This is because the authors feel that with a herd of cattle their interactions and contact patterns are relatively homogeneous within the herd. Thus it is not necessary to model those dynamics with an agent-based model

but in the between herd dynamics it is more important to model the heterogeneity that occurs in these interactions.

Agent based models are a way to capture the heterogeneity in a system that helps to drive the dynamics of that system. However, the heterogeneity can result in larger models that take more computational power and time. Hybrid models are a way to still capture that heterogeneity while reducing the computational power necessary to run the model. It is important though to decide which parts of the model the fidelity can be reduced in and made equation based or else the model will lose performance. While making the disease portion of the model equation based can save time and computing power it misses capturing individual agents actions and the importance of contacts and different contact patterns between agents in the spread of a disease. Switching between agent-based and equation based models allows for the contact patterns in the early stages of the outbreak to help drive the infectious disease spread but when the outbreak gets large enough for a few individual movements and interactions to no longer have as large of an impact on the outbreak because there are enough other agents infected. However, switching the entire model over or an entire city ignores the fact that transportation between cities and the movement of agents is not homogeneous.

We propose a model that differs from previous hybrid models in the literature by taking advantage of both versions of the hybrid architecture. The model uses a switching point to change from agent-based to equation-based models and back. We hypothesize that when enough agents are infected in a town the heterogeneous

mixing of agents within that town are not as important, so once a threshold of infected agents is reached switching to an equation based model will not result in a significant loss of fidelity. However, the entire model does not switch instead only the disease model switches to an equation-based model and the rest of the model remains agent-based. This allows the agents to still move in a realistic manner so that the spread of the disease throughout a network of towns still remains realistic.

10.2 Town Hybrid Model

We first test the hybrid model at the town level instead of the county level, as we want to make sure that the hybrid architecture works as expected on a smaller population before moving to the more time intensive county model. Our town hybrid model is created based on the models in Chapter 5 and 9. There are two main motivations behind the changes. The first is to improve efficiency of the model: both the environment component and the transportation component are altered to improve efficiency and to make the Chapter 5 more similar to the Chapter 9 model. The second change is to create a hybrid architecture: the disease component is altered to implement a hybrid model. The model is a hybrid agent-based and equation based model where the disease component of the model switches between agent-based and equation based when a certain percentage of agents are infected. The model is tested using the town of Schull in Ireland.

10.2.1 Model Components

The following sections give a brief overview of the model describing the four main components of an agent-based model taxonomy outlined in Chapter 4.

Environment Component

For the town hybrid model discussed in this section, the environment component is similar to that described in the county level model discussed in Section 9.1 but for a single town instead of a network of towns. Each small area is represented by one grid cell or Netlogo patch. When in a small area an agent has access to certain information such as the number of schools or workplaces in the small area and the real world distance to each other small area in the model. Furthermore, all the agents in a small area are physically in the same location, however, the agents keep track of where they are in that small area: home, work, school or community and restrict their interactions with other agents accordingly.

Society Component

The society is based off of real world census data from the Irish Central Statistics Office (CSO, 2014a) and is created in a similar fashion to how the society in the county level model was created. The society component for that model is described in Section 9.2. For each small area we create a population that reflects the population statistics of that small area including age, sex, household size and economic status. To be able to fully test the hybrid model any previous immunity

to the disease is not included in the model. This allows for larger outbreaks and more switching in the hybrid model. Social networks are included in the model. Agents have a family social network that is made up of any agents living in their household. Agents also have a work or school social network that is made up of other agents in their workplace or their school and students are given an additional social network which is a class network that is made up of agents who are in their school and of the same age. Social networks help to determine the contacts an agent has in the model.

Transportation Component

Transportation in the model varies from the Chapter 5 model. Instead of moving in steps between a location and desired destination agents move in one step. Agents movements are determined in one of two ways. Movements are either predetermined with the agents moving between home and school or home and work at certain times in the model or are determined randomly when an agent moves through the community. Agents moving through the community will pick a destination randomly from the small areas in the model. Although random movement is not completely realistic, at a small scale we feel that it is an acceptable approximation of how agents will move through a town.

Disease Component

The disease component of the model is made up of two different types of models: an agent-based disease component and an equation based disease component. It is

set up so that the model can be run with a completely agent-based disease model, a completely equation based disease model or switch between the two based on certain criteria. The following sections discuss the agent-based model, the equation based model and the method of switching between the two.

Disease Component: Agent-based Model The agent-based element of the disease component is based off of a compartmental Susceptible, Exposed, Infected, Recovered (SEIR) model. Where agents start in one of four different compartments, they are either susceptible, exposed, infected or recovered and based on their interactions they move between the compartments. The agent-based disease component remains unchanged from the Chapter 9 model. When an infectious agent comes into contact with a susceptible agent they have a given percent chance of passing the disease to the susceptible agent. If they do pass on the disease, the susceptible agent then moves to the exposed state for a given period of time before moving to the infectious state. They will remain in the infectious state for a set period of time before recovering. Once recovered they can not be reinfected. The disease dynamics are set to mimic measles.

Disease Component: Equation based Model The equation based part of the disease component uses an SEIR difference equation model. Difference equations were chosen over the more common differential equation models because of the discrete time space that are used in difference equations. This is more analogous to the agent-based model which also uses discrete time and will allow for

a more seamless transition between the two models. In the simulation, each geographic area selected runs its own SEIR difference equation model. The model can be run at the small area level or the town level. The equations used are similar to those discussed in Section 2.2, however, they are adjusted to include an exposed state. The equations are as follows:

$$S_{t+1} = S_t - \frac{\beta I_t S_t}{N} \tag{10.1}$$

$$E_{t+1} = E_t + \frac{\beta I_t S_t}{N} - \sigma E_t \tag{10.2}$$

$$I_{t+1} = I_t + \sigma E_t - \gamma I_t \tag{10.3}$$

$$R_{t+1} = R_t + \gamma I_t \tag{10.4}$$

Where S_t is the number of susceptible agents in the small area in the previous time step and S_{t+1} is the number of susceptible agents in the geographic area in the current time step. E_t and E_{t+1} are the number of exposed agents in the small area in the previous and current time steps, I_t and I_{t+1} are the number of infected agents in the small area in the previous and current time steps, and R_t and R_{t+1} are the number of recovered agents in the small area in the previous and current time steps. β is the infection rate or the probability of infection per contact between agents, σ is the rate of moving from exposed to infected and γ is

the recovery rate.

In a fully equation based disease component, each geographic area starts its difference equation model when an infected or exposed individual enters the area. In a hybrid model the difference equation model will start when the number of infected or exposed individuals is over a certain threshold. The threshold is discussed further in the next section. This could happen in two ways, either an agent from outside the area who is already exposed or infected moves into the area or an agent who is from the area becomes infected outside and returns home. Once the difference equation model has started it continues until there are no longer any more exposed or infected agents in the model.

At each time step, each area will calculate the values for the difference equations and adjust the number of agents in the area in each category. If the rounded difference between E_{t+1} and the number of exposed agents in the area is greater than 0, that number of susceptible agents in the small area will randomly be selected to move from the susceptible category to the exposed category. Similarly if the rounded difference between I_{t+1} and the count of infected agents in the area is greater than 0 than that number of exposed agents will be randomly selected to move from exposed to infected. If the rounded difference between R_{t+1} and the count of recovered agents in the area is greater than 0, than that number of infected agents in the area will recover.

Because movement is possible, there are times when the total number of agents in the area in one of the four categories is less or greater than the value predicted

in the model. Adjustment are made to account for this. If the value for E_t, I_t, or R_t is less than one and the count of agents exposed, infected or recovered in the area is greater than one then the value for E_t, I_t, or R_t is changed to the count of agents in that area who are exposed, infected or recovered. If the values for the difference between E_t, I_t, or R_t and the number of agents exposed, infected or recovered respectively in the geographic area is greater than the number of agents who could potentially move into the compartment (if the difference between E_t and the count of agents exposed is greater than the number of susceptible agents) the value for E_t, I_t, or R_t are adjusted down to reflect the actual counts of agents in the geographic area.

Switching

The model allows for geographic areas to switch between the equation based model and the agent-based model. The idea behind using a switch is that the agent-based models are especially important when a few agents are sick because at this stage the individual movements are what drive the spread of the disease so the heterogeneous movements of agents are more important. However, once the number of infected individuals reaches a certain number the individual movements should not matter as much because there are so many agents infected.

The decision of which model is used in a geographic area in a given time step is determined by the number of agents infected in that area. The user can set the switch threshold to be any percentage of agents infected or exposed and the

area will automatically switch between the agent-based model and the equation based model when this threshold is passed. Note, that if the number of infected or exposed agents in an area drops back below this threshold the model reverts back to an agent-based model.

In the town model we consider two levels of the switch, the small area level and the town level. If the switch occurs at the small area level then each small area will keep track of the number of agents who are infected and exposed in that small area. When the percent of agents who are exposed or infected in the small area is equal to or greater than the selected switch value the model switches from an agent-based disease component to an equation based disease component. Each small area will run its own set of difference equations and so the model can have some small areas running an agent-based disease component and some running an equation based component. When the percent of agents in the small area who are exposed or infected goes below the switch value then the small area returns to an agent-based disease component. If the switch is at the town level when the total percent of agents exposed or infected is greater than the switch value the whole model switches to an equation based disease component. When the percent of agents in the model who are exposed or infected is below the switch value the model returns to an agent-based disease component.

It is important to note that if the switch is set to 100% the model will be completely agent-based. If the switch is at 0% the disease component will always be equation based. If the switch is at the town level then a switch at 0% results in

an entirely equation based model as the location of a given agent does not influence if that agent becomes infected. This is because when the model is switched at the town level all agents are considered in the same equation based model and will thus mix homogeneously. Thus the results of the town hybrid model with a switch at 0% are only influenced by the initial conditions of the model such as the total number of agents, the number of initially infected agents or the number of immune agents at the start of the model.

10.2.2 Model Evaluation

In this section we report a number of experiments on our hybrid town model that were designed to test whether our hybrid model successfully blends the fidelity of agent-based models with the computational efficiency of equation based models. We ran a number of experiments to evaluate the performance of our hybrid model at the town level. In these experiments we treat the behaviour of a completely agent-based model (i.e., a model with a switch threshold of 100%) as the ground truth because agent-based models are considered to have the higher fidelity of the two modelling approaches. Comparing the results of an hybrid model to an agent-based model is used in the literature with Bobashev et al. (2007) using their agent-based model as the standard to compare their hybrid to as it has the most micro level detail. Consequently, if the hybrid model produces similar results to a completely agent-based model, while using less computational resources, then we can consider our hybrid modelling approach to be successful.

Note, that there are two hyper-parameters that may affect the performance of the hybrid model. The first hyper-parameter is the geographical scale that the switch is applied at: small-area or town level. The second is threshold within the relevant geographic area that is used to switch between the agent-based and equation based models. To test the interactions between these hyper-parameters and the hybrid model performance, in each of the following experiments we run the following hybrid models: town switch with 0% threshold, town switch with 10% threshold, town switch with 20% threshold, small-area switch with 0% threshold, small-area switch with 10% threshold, small-area switch with 20% threshold, small-area switch with 30% threshold, and small-area switch with 35% threshold. Also, because of the stochastic nature of agent-based models we run each model 300 times and use statistics calculated across these runs to compare with other models.

Within the above experimental framework, the first experiment we report is a sense-check analysis that counted the number of switches a hybrid model makes between the agent-based and equation based component. The motivation for this experiment was that if we found that a hybrid model rarely switches, and remains agent-based for the majority of the runs, then the hybrid model is not useful. The second experiment we report analyses the time-saved by a hybrid model when it switches to an equation based disease component. To examine the time saved we compare the average number of seconds needed per time step of the hybrid model with a fully agent-based model. The final two experiments we report in this section are designed to compare the fidelity of the hybrid models with the fully

agent-based model. The first of these fidelity experiments analyses the divergence between the number of infected agents in the hybrid models and the fully agent-based model. The second fidelity experiment analyses the divergence between the length of outbreaks in the hybrid models and the fully agent-based model.

Finally, switching to the equation based disease component in the hybrid architecture will result in a loss of fidelity in the model results as the advantages of the agent-based disease component are lost. However, some of the advantages of using the equation-based component might out weight the cost of losing the fidelity of the model. Consequently, we conclude these experiments by identifying a set of hyper-parameters (geographic switch area, and switch threshold) for our hybrid model that usefully balances between model fidelity and time savings.

Number of Runs that Switch

We first look at the number of the 300 runs that results in the disease model switching to equation based. There are some cases where the model does not switch over to the equation based because the required number of agents are never infected. Table 10.1 shows the percentage of runs that the model switches for the small area switch and the town switch along with the 95% confidence intervals for each value.

For both the town switch and the small area switch models for the switch values we look at the model switches to an equation based model in a majority of the runs. This, however, only measures if a model switches over for at least

	Percent of Runs that Switch	
Switch	Small Area	Town
10%	93.0	92.7
	(90.1, 95.9)	*(89.7, 95.6)*
20%	93.7	83.0
	(90.9, 96.4)	*(78.7, 87.3)*
30%	92.0	-
	(88.9, 95.1)	-
35%	85.0	-
	(81.0, 89.0)	-

Table 10.1: Percent of runs that lead to the model switching from agent-based to equation based for different versions of the hybrid model

one time step. While it is a good starting point to look at if the disease model is actually switching between agent-based and equation based, the length of time a model switches for should also be studied. To do this we find the number of time steps the model is using an equation-based disease model. The distributions for the number of time steps the model switches for the small area switch model and the town switch model can be seen in Figures 10.1 and 10.2 respectively.

For both models it can be seen that as the switching percentage increases, the number of time steps that use an equation based model decreases. This is as expected as a higher percentage equates to a larger number of agents required to be exposed or infected before the model switches to equation based from agent-based and infecting a larger number of agents will take more time steps.

Run Time

It is also important to determine if using the hybrid will actually results in real savings when running the model. To test for this we find the average number of seconds per time step in each of our versions of the model. Table 10.2 shows the

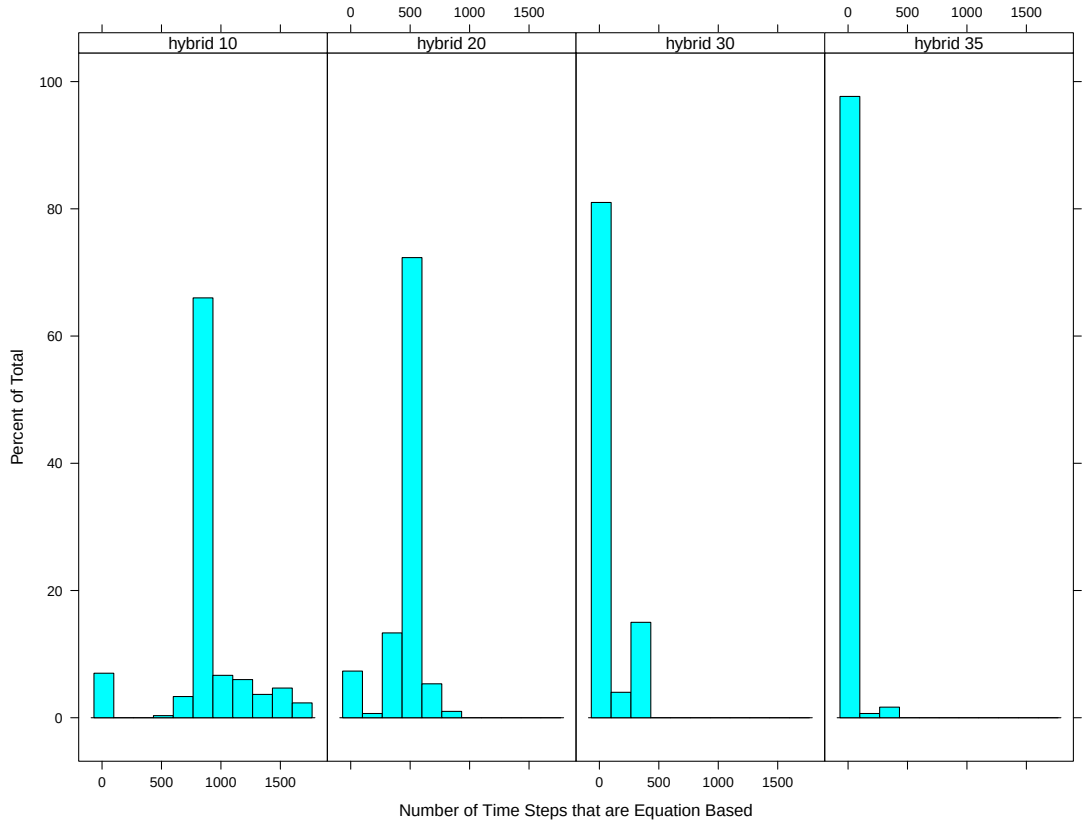


Figure 10.1: Distribution of the total number of time steps the disease model is equation based for the small area switch model. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based.

times for the model with six different switching scenarios: where the disease model is always equation based, where the model switches to equation based when 10% of agents are infected or exposed, where the model switches to equation based when 20% of agents are infected or exposed, where the model switches to equation based when 30% of agents are infected or exposed, where the model switches to equation based when 35% of agents are infected or exposed, and where the disease model is always agent-based. The table also provides a 95% confidence interval for the average times. At 30% the town model no longer switches from agent-based

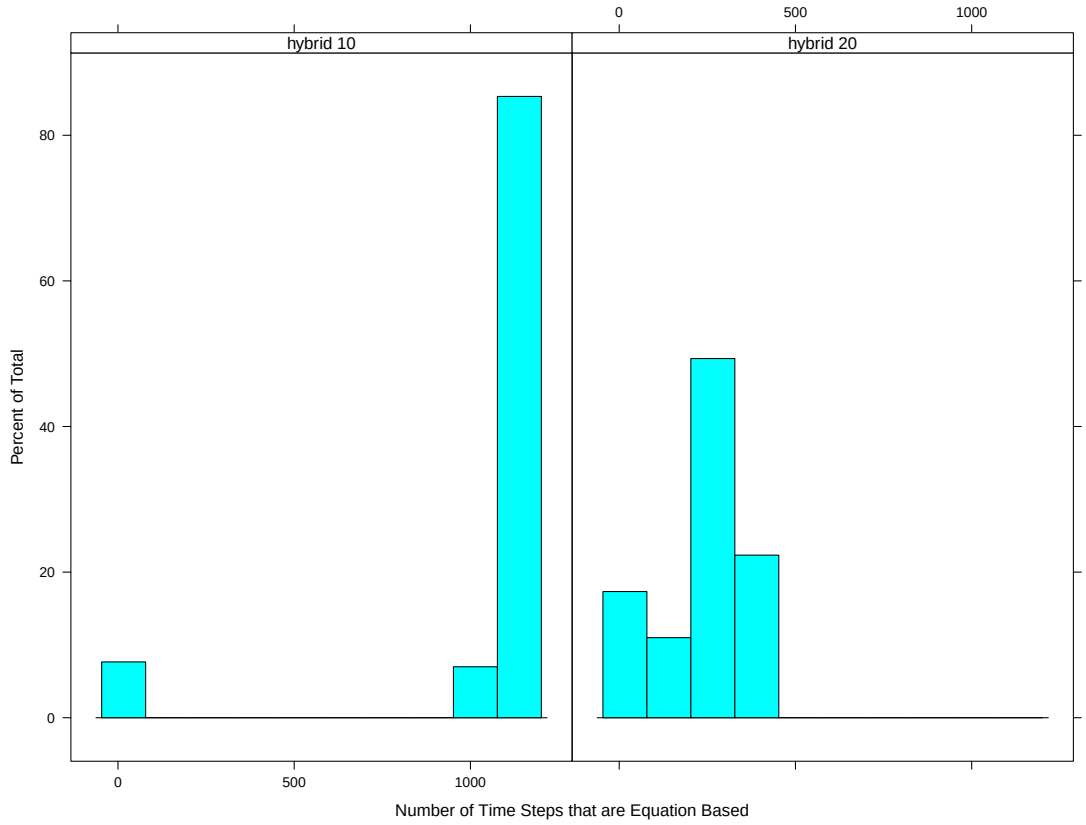


Figure 10.2: Distribution of the total number of time steps the disease model is equation based for the town switch model. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based.

to equation based so the values are not shown in the table. At 40% the small area switch model switches in less than half of the runs and does not result in any time savings so the results for any switches after 35% are not included in the table for either version of the model.

From the table it can be seen that for the two switching versions, the small area model results in more time saved per time step when compared to the fully agent-based model. In both cases switching over at 10% provides greater savings than switching at 20%. This makes sense as it takes significantly less time per

	time (ms) per time step	
Switch	Small Area	Town
0% (Equation Based Disease Component)	1.79	1.90
	(1.59, 1.99)	*(1.68, 2.11)*
10%	2.42	3.55
	(2.14, 2.69)	*(3.15, 3.95)*
20%	4.05	4.61
	(3.59, 4.50)	*(4.08, 5.13)*
30%	5.23	-
	(4.68, 5.87)	-
35%	5.77	-
	(5.12, 6.43)	-
100% (Fully Agent Based Model)	6.77	6.77
	(6.01, 7.54)	*(6.01, 7.54)*

Table 10.2: Average number of milliseconds for a time step for different versions of the hybrid model switching at the small area level

step when the disease model is completely equation based versus agent-based, therefore, the longer the model stays at agent-based before switching to equation based the longer the average step length will be. This can be further seen in that the time per step increases in the small area model when the switch is at 30% and 35%. However, the time per step for these two switch points are not significantly different from each other as their values fall with in the others confidence intervals. These results are taken as evidence that the hybrid model is successfully providing time savings as compared with the pure agent-based model.

Distribution of Number Infected

After looking at the time saved and the switching behaviours of the model the results are analysed. For each of the different versions of the hybrid model the results are compared to the completely agent-based version of the model. Figure 10.3 shows the distribution of the number of infected agents across the 300 runs

for the small area model at different switch values.

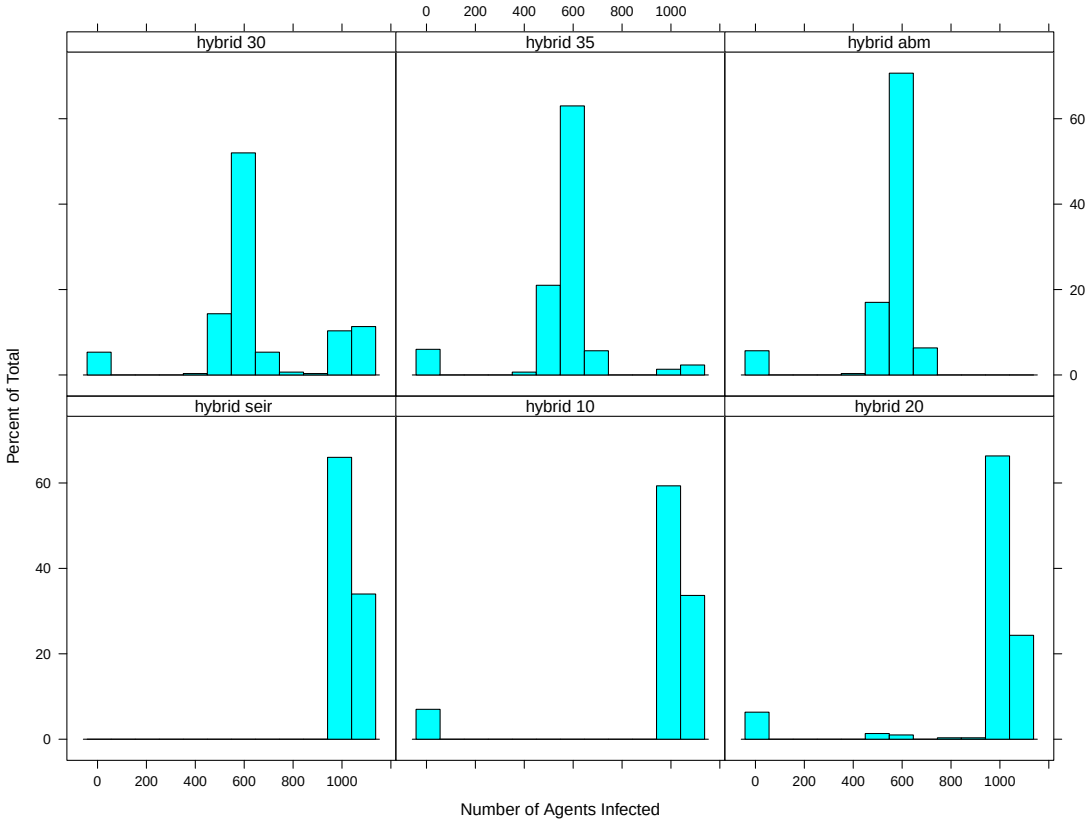


Figure 10.3: Distribution of the total number of agents infected by run for the small area switch model

It can be seen in the figure that as the switch gets higher, a higher percent of agents need to be exposed or infected before the model switches to an equation based model, the distribution moves away from the version of the model where the disease model is strictly equation based and moves towards the strictly agent-based version. A similar pattern is seen in the town model. The distributions for the model switching at the town level can be seen in Figure 10.4.

Although it is useful to visualize the change in distribution, it is possible to actually compare the distributions and get a value for the probability that the

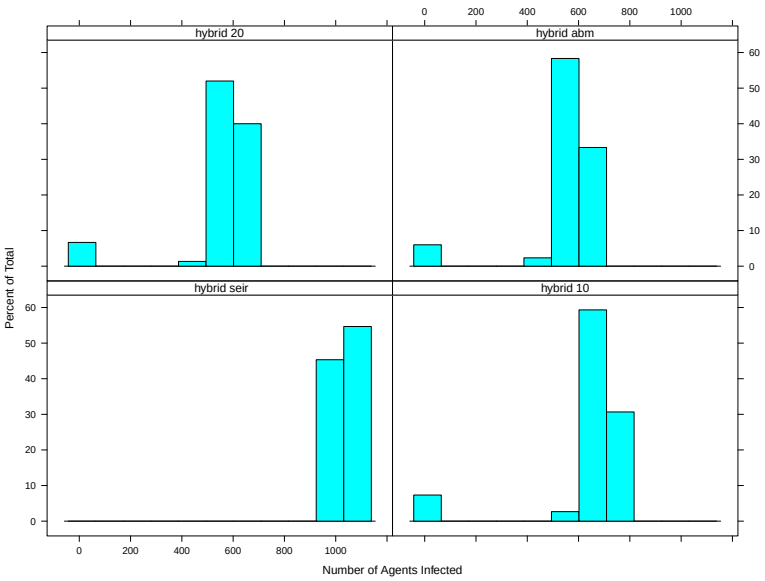


Figure 10.4: Distribution of the total number of agents infected by run for the town switch model

sample distributions come from the same population. The Wilcoxon rank sum test is a non parametric alternative to a two-sample t-test that does not assume the population distribution is normal. The null hypothesis of the test is that the two populations have the same distribution. A Wilcoxon rank sum test is done for each of our distributions from the switching models compared to the distribution from the completely agent-based model. The p-values for those tests can be found in Table 10.3.

	P-value	
Switch	Small Area	Town
10%	0.000	0.000
20%	0.000	0.015
30%	$5.9e^{-6}$	-
35%	0.5791	-

Table 10.3: P-values for the Wilcoxon rank sum test comparing the outbreak size distributions for the switching models to the completely agent-based model.

The values show that as the switch threshold is higher, the distribution gets

closer to the agent-based model. This can easily be explained, the larger the switch the longer the model remains agent-based so the more similar the two distributions will be. Our aim was to find a range of switch point that still results in the model switching between an agent-based and equation based disease model but also results in a distribution that is similar to the complete agent-based model. From the table we can see that a switch of 35% at the small area level results in a distribution that is not significantly different from an agent-based model and that a switch of 20% at the town level results in a distribution that is not significantly different from the agent-based model distribution at a 1% significance level.

Length of Outbreak

Finally the total time it takes for an outbreak to finish is investigated. An outbreak is considered finished if there are no agents exposed or infected within the model. The outbreak length is studied because it is another important characteristic of model output. If the hybrid model has a similar outbreak size but a different outbreak length than its not possible to say that the outbreaks are similar. The distributions of the number of time steps it takes for the outbreak to finish for the small area switching model and the town switching model can be seen in Figure 10.5 and Figure 10.6 respectively.

Similar to the total outbreak size distributions the outbreak length distributions converge to the completely agent-based model distribution. Again the Wilcoxon rank sum test is used to compare the time distributions. The values can be found

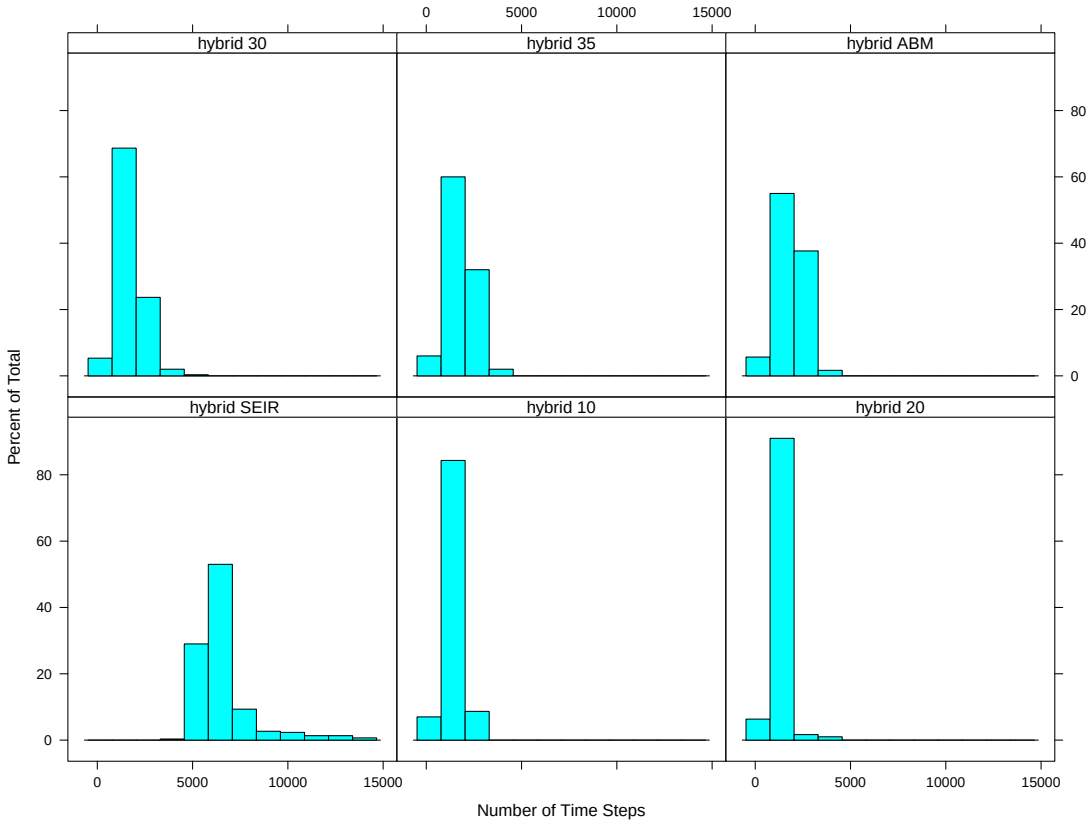


Figure 10.5: Distribution of the total number of time steps for the small area switch model to finish

in Table 10.4. From the table it can be seen that as the switch threshold increases the distribution gets closer to that of the agent-based model. For the switch at the small area level we can see that when the switch is 35% the distributions are not significantly different at 10% significance level and for the switch at the town level when the switch is at 20% the distributions are not significantly different at the 5% level.

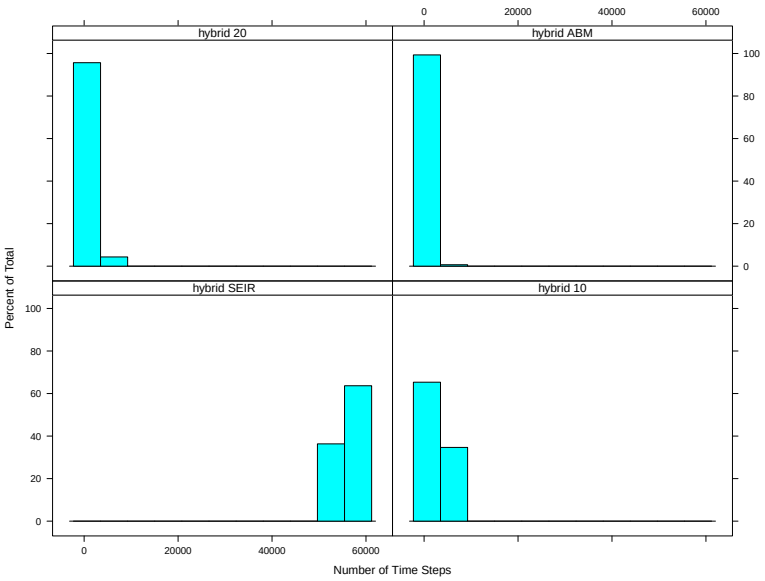


Figure 10.6: Distribution of the total number of time steps for the town switch model to finish

	P-value	
Switch	Small Area	Town
10%	0.000	0.000
20%	0.000	0.0707
30%	$1.6e^{-7}$	-
35%	0.1036	-

Table 10.4: P-values for the Wilcoxon rank sum test comparing the outbreak time distributions for the switching models to the completely agent-based model.

Discussion

Based on the above results we can see that the hybrid model is able to switch between agent-based and equation based for a majority of runs when the switch is 35% or below and switching at the small area level and 20% or below and switching at the town level. In addition, at all values of the switch considered we see that there are significant time savings when running the hybrid model over the purely agent-based model. The experiments to analyse the fidelity of the hybrid model show that at lower switch values, the hybrid model distributions

for both the total number of agents infected and the length of the outbreak are significantly different from the agent-based model distributions. However, at a switch of 35% for the small area switch and 20% for the town switch statistical tests show that there is not a significant difference between the hybrid and agent-based distributions. Although both the small area and town level models result in time savings and produce significant results, the time savings are greater at the town level as the town level switch converges faster to the agent-based model results. A hybrid model switching at the small area level with a threshold switch of 35% is statistically similar to a fully agent-based model and has a time savings of an average of one millisecond per time step while a hybrid model switching at the town level with a threshold of 20% is also statistically similar to a fully agent-based model but has a time savings of an average of 2.16 milliseconds per time step. Because of this we feel that switching at a town level over a small area level provides a greater advantage.

10.3 County Hybrid Model

The hybrid model for a single town is a start in an analysis to show that a hybrid model can succeed in both saving time and computing power when running a large agent-based model. The results show that not only does a hybrid model save computing time compared to a fully agent-based model but the results also start to converge to the results for the agent-based model as the switch point changes. However, even though in most cases the models appear to be converging the results

are still shown to be from different distributions based on the Wilcoxon rank sum test and any larger switch values will not save time or result in the model actually switching. One factor causing this could be that the model is run on a small town. With only about 1,000 agents in the entire model switching can only happen on a small scale. In addition, at such a small scale the fully agent-based model does not take too much time to run leading to advantages in time saved for the hybrid model being negligible in many cases. To show the true advantage of a hybrid model it will be necessary to start with a model that is much larger where saving time can be done on a larger scale. To do this a county model is used. The county model is a scaled up version of the town model in the previous section.

10.3.1 Model Components

The county hybrid model is identical to the county model in Chapter 9 except for the disease component allows for the model to switch between agent-based and equation based. Both the agent-based and equation based disease components of the model are the same as that of the town model. Switching is also similar, however, the county model switches at either the town level or the whole county level. The model being used in a given time step by the town is determined by the number of agents infected in either the town or the whole county. Similar to in the town model, it is important to note that if the switch is 0%, the model will switch from agent-based to equation based when 0% of agents are exposed or infected, this means the disease component of the model will always by equation based and

if the switch is at 100% the model will always be completely agent-based. However, when the disease model switches at the county level, there is one set of difference equations for the whole county, the model is essentially completely equation based. Even though agents are allowed to move, because agents are infected at the county level their location does not have an influence on if the agent will be infected or not. The model with a switch of 0% has no stochasticity in it and the only thing that would have an impact on the results is the initial conditions: if the model starts with more or less agents, more than one agent infected, or there are a number of agents who are already immune.

10.3.2 Model Evaluation

To test the county hybrid model we run similar experiments to those presented in Section 10.2.2 to look at the switching behaviour of the model, the time savings and the fidelity of the results when compared to the fully agent-based model. We do one additional fidelity test for the hybrid model to compare how the outbreak spreads through the network of towns in the county.

For both the town switch and the county switch we look at the switch values of 0%, 5% and 10% and we also look at switches at 20% and 30% at the town level. For each switch value except for 0% we run the model 300 times. As mentioned in the previous section there is no stochasticity in the model with a switch of 0% at the county level thus the model only needs to be run once to get the results.

Number of Runs that Switch

In order to make sure the model is utilizing the hybrid architecture we look at a number of measures: the number of time steps that the model has switched to hybrid and the maximum number of towns that have switched to hybrid during the model.

Table 10.5 shows the percent of runs for each of the switch values that results in the model switching to an equation-based disease component for at least one time step.

	Percent of Runs that Switch	
Switch	Town	County
5%	76.3	89.7
	(71.5, 81.1)	*(86.2,93.1)*
10%	73.3	88.0
	(68.3, 78.3)	*(84.3, 91.7)*
20%	75.3	-
	(70.5, 80.2)	-
30%	69.3	-
	(64.1, 74.6)	-

Table 10.5: Percent of runs that lead to the model switching from agent-based to equation based

The table shows that for all versions of the switch the model becomes equation based for a large portion of runs. It can also be seen from the model that at the 5% and 10% threshold it is more likely for a switch to occur if the model is switching at the county level versus the town level. However, if the switch value is 20% the model switching at the county level will not switch to equation based.

Showing if the model switches at all is important but it is also important to look at the number of time steps that are equation based in the model. A run is

counted in the percentages in Table 10.5 if it has switched to equation based for at least one time step, but models switching for only one or two time steps are not taking full advantage of the hybrid architecture of the model. Thus we look at the distributions of the number of time steps that have switched from agent-based to equation based. Figure 10.7 shows the distribution when switching at the town level and Figure 10.8 shows the distribution when switching at the county level.

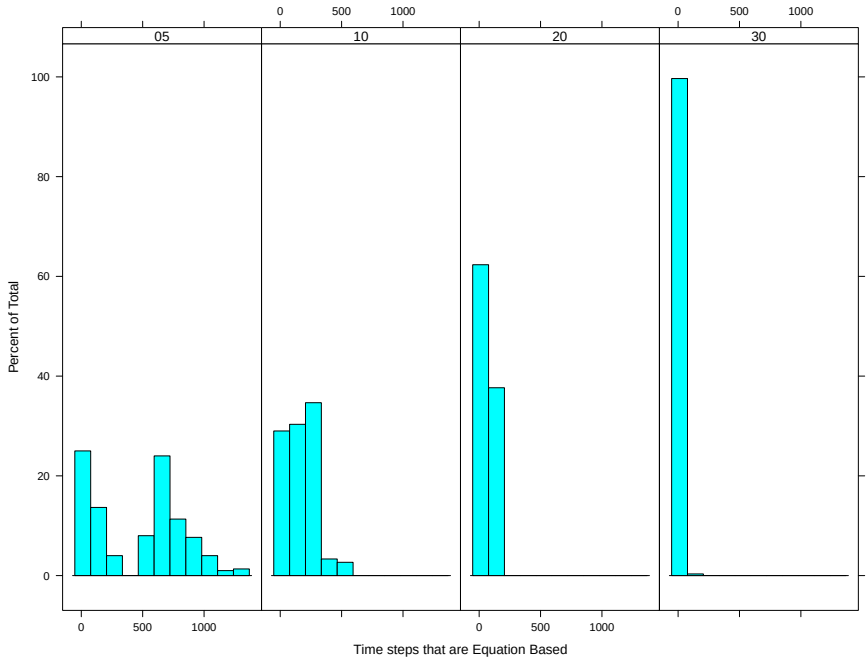


Figure 10.7: Distribution of the total number of time steps the disease model is equation based when the model switches at the town level. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based.

Looking at the distributions of the count of time steps where the model has switched to an equation based disease model it can be seen that when the switch value is lower, the number of time steps where at least one town has switched to equation based increases. This is as expected and makes sense as the model should

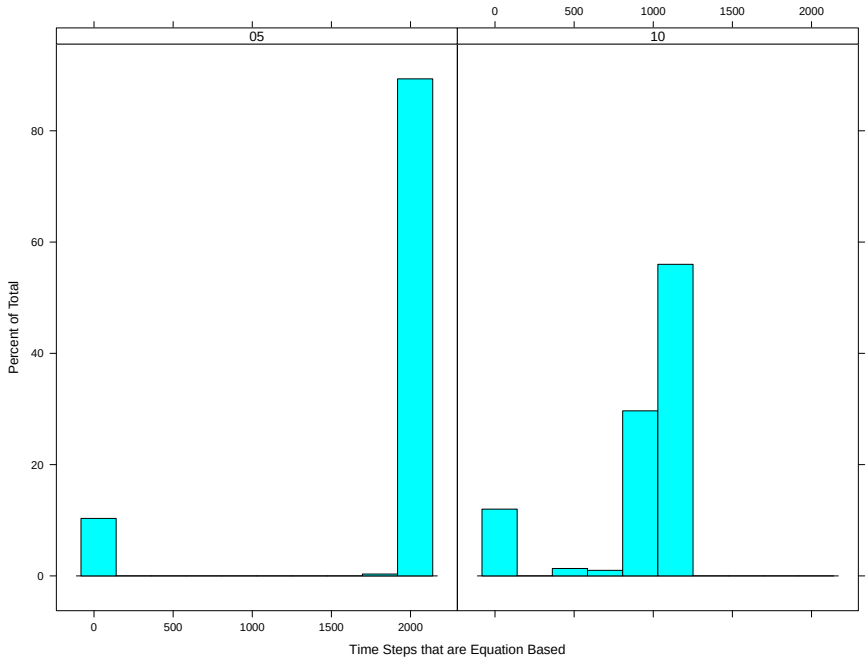


Figure 10.8: Distribution of the total number of time steps the disease model is equation based when the model switches at the county level. The zero column in the histograms reflect the number of runs where the model does not switch from agent-based to equation based.

reach the point where 5% of agents are infected or exposed before 30% of agents are infected or exposed and thus will remain equation based for longer.

The maximum number of small areas that have had their town switch to an equation based model can be found in Figure 10.9. This is only done for the town switch model because when the model switches at the county level all towns switch together at the same time. As expected the model with a lower switch value has a higher maximum number of small areas that have switched to equation based. The town switch model does not result in a larger portion of the model switching at any one time. With a switch of 5% the maximum number of small areas switched is 16 out of a total of 173 small areas in the county. This number reduces even

more as the switch increases to 30% with only a maximum of 4 small areas in the equation based model at any given time step.

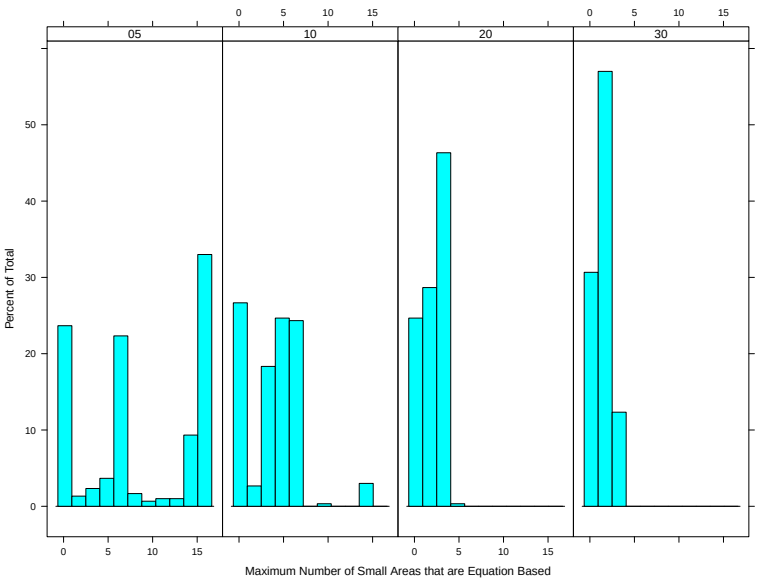


Figure 10.9: Distribution of the maximum number of small areas that have an equation based disease component at a given time

Run Time

Similar to the run time experiment in Section 10.2.2 we look at the run time of the model to determine if the hybrid model produces savings over the fully agent-based model. Table 10.6 shows the results for the average time in seconds for each time step in the model.

From the table it can be seen that there is almost half a second time savings per step going from a full agent-based model to a model where the disease component switches to equation based when 5% of the agents are infected or exposed. Additionally we can see that when the model switches when there are 20% or 30% of agents infected or exposed, there are not significant time savings when compared

	time (s) per step	
Switch	Town	County
0% (Equation Based Disease Component)	0.82	0.64
	(0.66, 0.98)	*(0.52, 0.77)*
5%	1.00	0.87
	(0.80, 1.20)	*(0.70, 1.04)*
10%	1.04	1.21
	(0.84, 1.25)	*(0.97, 1.44)*
20%	1.28	-
	(1.03, 1.53)	-
30%	1.36	-
	(1.09, 1.63)	-
100% (Fully Agent Based Model)	1.40	1.40
	(1.12, 1.67)	*(1.12, 1.67)*

Table 10.6: Average number of seconds for a time step for different versions of the hybrid model

to the completely agent-based model.

A similar time savings of over a half a minute can be seen in the model that switches at the county level going from the agent-based model to the hybrid model that switches at 5% infected and exposed. However, even though the average number of seconds per time step is 0.2 seconds less than the agent-based model when the switch is 10%, the average value falls within the confidence interval of the agent-based model and vice versa showing no significant difference.

Distribution of Number Infected

Determining the time savings and the switching behaviour of the the model allows us to determine if the hybrid architecture is both working by saving time and resulting in the model switching between agent-based and equation based for an extended period of time. Once it is determined that the model is working, it is necessary to look at the results and look at how the results of the hybrid model

compare to the complete agent-based model. The distribution of the total number of infected agents across the runs for the different switching points can be found in Figure 10.10.

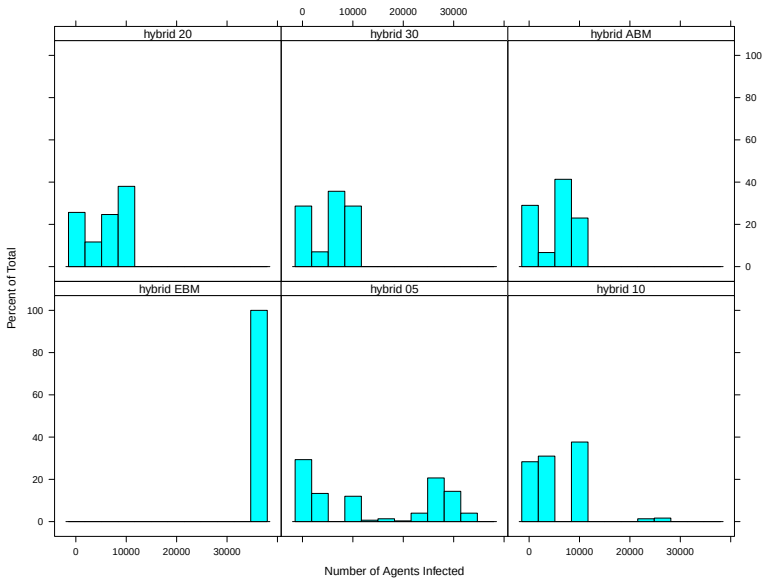


Figure 10.10: Distribution of the total number of infected agents when the county model switches at the town level.

From the figure it can be seen that the distributions for switching at 20 or 30% infected or exposed are very similar to the fully agent-based model. Switching at 10% infected or exposed still results in a similar distribution, however there appear to be some more obvious differences such as a small cluster of outliers to the right of the distribution representing a number of runs with a much higher number of total infected agents. It can also be noted that comparing the 10% switching model to the fully agent-based model that there is a higher number of runs with a smaller number of infectious agents when the model switches. The distribution for the 5% switching model looks distinctly different from the rest of the models.

The 5% switching model results in a distribution with a much larger number of agents infected then any of the other models. All of the switching hybrid models are different from the model where the disease component is fully equation based, however it can be seen how the model results move farther from the equation based disease component model as the switch increases. The differences in the equation based model, 5%, 10%, and full agent-based model help to show why agent-based models are important. When homogeneous mixing is present in the model, the equation based disease model, a higher number of agents can become infected. This is because all agents have equal probabilities of coming into contact with each other if they are in the same town. However, when heterogeneous mixing is used in the model (the agent-based disease model) there are fewer infected agents because agents are less likely to infect someone outside of their social networks.

A similar analysis is done when the switch is at the county level. The distribution of the total number of infected agents can be found in Figure 10.11. From the figure it can seen that the models that switch from an agent-based to an equation-based disease component appear more similar to the model with an equation based disease component then an agent-based disease component. They do, however, appear to be slowly converging towards the agent-based results.

To further compare the similarities of the distributions, the Wilcoxon signed-rank test is run comparing the hybrid models to the completely agent-based model. The tests are used to determine if two sample distributions come from the same population. Table 10.7 shows the p-values for the tests comparing each hybrid

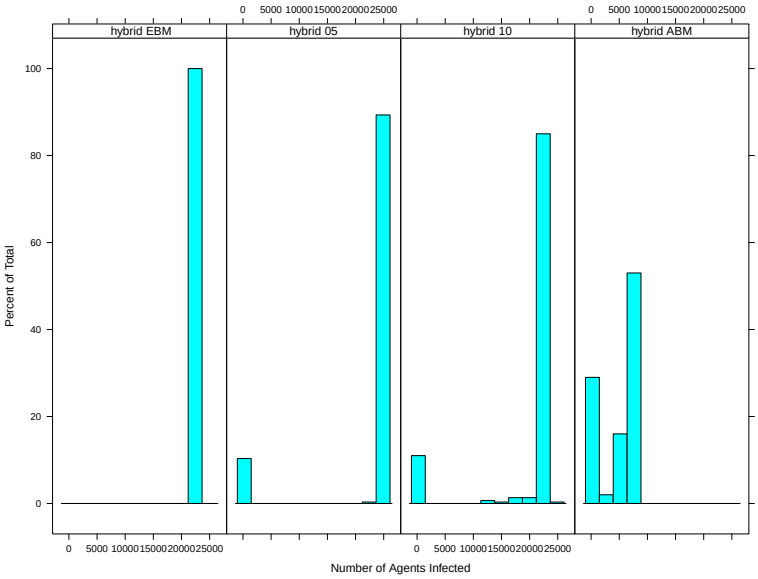


Figure 10.11: Distribution of the total number of infected agents when the county model switches at the county level.

models to the fully agent-based model.

	P-value	
Switch	Town	County
5%	$4.861e^{-12}$	$2.2e^{-16}$
10%	0.1625	$2.2e^{-16}$
20%	0.2942	-
30%	0.9566	-

Table 10.7: P-values for the Wilcoxon rank sum test comparing the outbreak size distributions for the switching models to the completely agent-based model.

The p-values further show what was seen Figure 10.10. For the model that switches at the county level, the p-values are close to 0 meaning that the null hypothesis of the distributions coming from the same population should be rejected. Thus switching at the county level does not result in distributions of infected agents that are similar to the agent-based model. When the switch is at the town level, the distribution with a 5% switch value has a p-value very close to 0 so the null should be rejected as well. However, the distributions for 10%, 20%, and 30%

are not significantly different from the agent-based model.

Length of Outbreak

To compare our outbreaks we also look at the time it takes for the outbreak to finish. An outbreak is complete when there are no longer any exposed or infected agents in the model. If the run times of the models are drastically different it will be hard to compare the results as the outbreak length is a key descriptive feature of an outbreak. To compare the lengths of outbreaks across models the distributions of the number of time steps taken for the model to finish is looked at for both versions of the model and each switch value. The distribution for the model that switches at the town level can be seen in Figure 10.12.

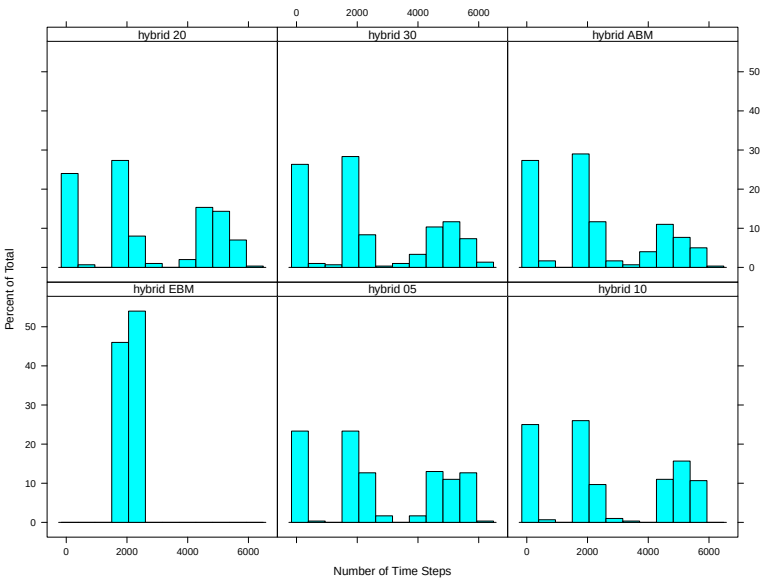


Figure 10.12: Distribution of the length of time taken for the model to finish when switching at the town level.

A similar analysis is done for the model that switches at the county level. The distribution of the number of time steps that it takes for the model to finish for

each of the four different versions of the model is found in Figure 10.13. Similar to the distribution of number of agents infected it can be seen that the number of time steps it takes for the model with a completely equation based disease component is distinctly different then the fully hybrid model. The two versions of the model that switch between the agent-based and equation based disease components have a distribution of time steps in between the equation and agent-based versions.

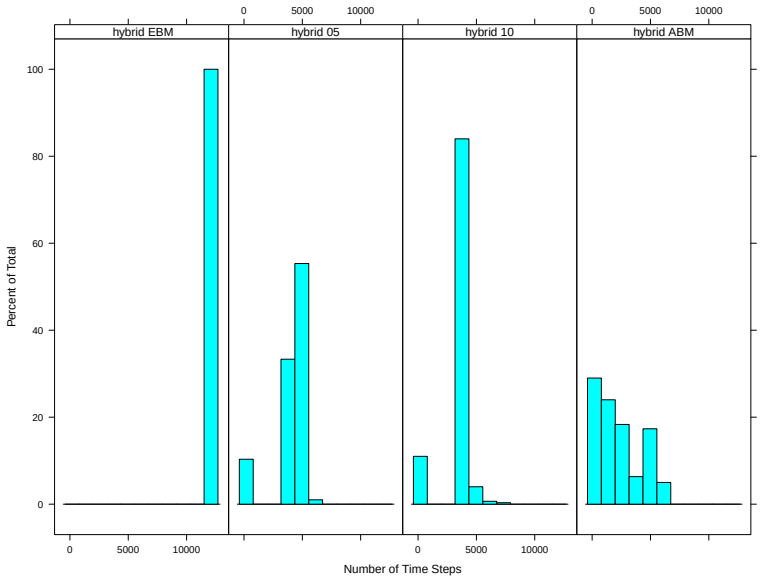


Figure 10.13: Distribution of the length of time taken for the model that switches at the county level to finish

The p-values for the Wilcoxon tests to compare the hybrid models to the agent-based model can be seen in Table 10.8, the p-values for the models that switch at 5%, 10% and 20% are all small meaning there is only a small probability that the distributions are from the same population. The distributions for the number infected that are from the model that switch at the county level show clear differences between the complete agent-based and the hybrid models. This can be

further seen in the p-values from the Wilcoxon tests that are very near zero.

	P-value	
Switch	Town	County
5%	0.0021	$2.2e^{-16}$
10%	0.0143	$2.2e^{-16}$
20%	0.0160	-
30%	0.2283	-

Table 10.8: P-values for the Wilcoxon rank sum test comparing the outbreak length distributions for the switching models to the completely agent-based model.

Spread of the Outbreak

Another aspect of the outbreak that can be considered is what towns the outbreak spreads to. As the model is run in a scenario with a highly infectious disease and the population has no previous immunity, there is a large number of infected agents in the model

Table 10.9 shows the percent of runs that lead to an outbreak in the switch model for twelve different towns in Leitrim County along with the population of the town and a weighted degree centrality. The degree centrality is a measure of the number of agents that commute in and out of the town. Six of the towns are larger towns that are made up of multiple small areas, Ballinamore, Dromahair, Leitrim, Lurganboy, Manorhamilton, and Mohill. The other six towns are smaller towns that are only made up of one small area, Aghacashel, Corrala, Glenfarn, Newtowngore, Munakill and Rinn. The results are given for each version of the model based on the switch and the fully agent-based model. The version of the model with the fully equation based disease component is not included as the model results in nearly all the agents becoming infected every run. Thus towns would

have an outbreak in it each run. From the results it can be seen that for the larger towns, the majority of the runs result in outbreaks and for the completely agent-based model all of the larger towns have 71% of runs that lead to an outbreak. This is the same percent as the percent of runs that leads to an outbreak in the overall model, with 71% of runs having at least two agents in the county infected. This is likely because the model is run without immunity to be able to fully test out the hybrid model and therefore for a larger town that is likely more central with more agents commuting in and out when there is an outbreak in the county where no agents are immune it will spread to the larger towns. As the smaller towns have fewer agents they are not as likely to have the outbreak spread to them.

Town	Centrality	Population	5%	10%	20%	30%	Agent-Based
Aghacashel	0.05	73	33.3	16.3	15.7	12.3	10.0
			(28.0, 38.7)	*(12.2, 20.5)*	*(11.6, 19.8)*	*(8.6,16.0)*	*(6.6, 13.4)*
Ballinamore	0.44	1096	76.3	74.0	75.3	72.0	71.0
			(71.5, 81.1)	*(6.09.0, 79.0)*	*(70.5, 80.2)*	*(66.9, 77.1)*	*(65.9, 76.1)*
Corrala	0.14	248	64.0	56.3	56.3	58	51.3
			(56.3, 69.4)	*(50.7, 61.9)*	*(50.7, 61.9)*	*(52.4, 63.6)*	*(45.7, 57.0)*
Dromahair	0.15	1506	76.3	74.3	75.3	72.3	71.0
			(71.4, 81.1)	*(69.4,79.3)*	*(70.5, 80.2)*	*(67.3, 77.4)*	*(65.9, 76.1)*
Glenfarn	0.03	147	52.7	35.0	32.3	28.7	27.7
			(47.0, 58.3)	*(29.6, 40.4)*	*(27.0, 37.6)*	*(23.4, 33.8)*	*(22.6, 32.7)*
Leitrim	0.21	1123	76.3	74.3	75.3	72	71.0
			(71.5, 81.1)	*(69.4, 79.3)*	*(70.5, 80.2)*	*(66.9, 77.1)*	*(65.9,76.1)*
Lurganboy	0.00	388	76.3	74.3	75.3	72	71.0
			(71.5, 81.1)	*(69.4, 79.3)*	*(70.5, 80.2)*	*(66.9, 77.1)*	*(65.9, 76.1)*
Manorhamilton	1.00	1782	76.3	74.3	75.3	72.3	71.00
			(71.5, 81.1)	*(69.4, 79.3)*	*(70.5, 80.2)*	*(67.3, 77.4)*	*(65.9, 76.1)*
Mohill	0.86	1378	76.3	74.0	75.3	72.0	71.0
			(71.5, 81.1)	*(69.0, 79.0)*	*(70.4, 80.2)*	*(66.9, 77.1)*	*(65.9, 76.1)*
Munakill	0.12	196	62.3	55.0	57.3	56.0	45.7
			(56.9, 67.8)	*(49.4, 60.6)*	*(51.7, 62.9)*	*(50.4, 61.6)*	*(40.0, 51.3)*
Newtowngore	0.14	230	62.0	45.3	42.7	38.3	39.3
			(56.5, 67.5)	*(39.7, 51.0)*	*(37.1, 48.3)*	*(32.8, 43.8)*	*(33.8, 44.9)*
Rinn	0.36	293	66.7	61.3	68.7	64.7	58.7
			(61.3, 72.0)	*(55.8, 66.8)*	*(63.4, 73.9)*	*(59.3, 70.1)*	*(53.1, 65.2)*

Table 10.9: Percent of runs that the outbreak spreads to the given town

It can also be seen from Table 10.9 that as the model gets closer to the fully agent-based model, the switch increases in size, the results are more similar to the fully agent-based model, further showing how the hybrid model converges to the agent-based model as the switch increases. The percent of runs that lead to an

outbreak is also calculated for the model with the county level switch. However, when the model that switches at the county level switches to an equation based disease component from an agent-based disease component the location of an agent does not have an influence on if they become infected and there is homogeneous mixing for all agents in the model. Thus the percent of runs that lead to an outbreak for the model switching at the county level when 5% of agents are infected or exposed and when 10% of the agents are infected or exposed is equal to the number of runs that spread outside of a single town in model. This shows a clear difference between the model that switches at the town level and the model that switches at the county level. The town switch still allows for agents movement patterns, such as their commutes to influence the spread of the outbreak. Even though the larger more central towns in the model have similar percent of runs that lead to an outbreak, the smaller less central towns have more variable results. This is not the case in the model that switches at the county level where the outbreak is equally likely to spread to all towns regardless of the size and centrality.

10.4 Conclusion

The chapter shows that it is possible to create a hybrid model for infectious disease epidemiology where the disease component switches between agent-based and equation based determined by the number of agents infected. We looked at a number of levels for switching, both at the actual value of the switch (5%, 10% etc.) and at the size of the area where the switch occurs (small area, town or

county). For each version of the hybrid model we compared the results to the fully agent-based model and found that a number of factors influence the results of the hybrid model. The value of the switch, if the model turns to equation based at 5% or 30% infected is important as it determines the initial conditions of the equation based component of the model. The higher the switch the less likely it will be that the model switches and switches for an extended period of time. The higher switch values do not result in as much savings of time and computing power as the lower switch values. In addition, these models were run on a scenario where the entire population was susceptible to the disease. While this may be the case for new and emerging diseases, for a disease such as or influenza a portion of the population will be already immune to the disease. This will create even less opportunity for switching at a higher percent of agents infected or exposed.

Another factor influencing the results of the model is the area over which the switch occurs. From our test we have looked at a number of levels from small area to town to county. The results of our model show that the smaller the area of the switch the less time saved, this is because at the lower levels the equation based and agent-based disease components will be running simultaneously based on the number infected at each town so more of the model will be agent-based even when the model has switched. However, at the county level the entire model will be equation based at once so there is more time savings. In addition, the size of the area that is switched has an impact on how similar the results will get to the agent-based model and the largest switch value that can be used. The smaller

the area that is switching the larger the switch can be. For example, when we switched at the small area level the model still switches at values up to 35% but the county model only switches to about 10%. We can also see that when the switch is at the town level the hybrid model converges to the agent-based model faster than if the the switch is at the small area level or the county level.

Our analysis leads us to the conclusion that at both levels of the model, town and county, the switch for our hybrid model is best done at the town level. We think that the town level switch provides sufficient time savings compared to a fully agent-based model while still being able to produce results that are similar to the fully agent-based model. Not only do we capture a similar distribution of the number of infected agents but the model is also able to capture a similar spread of the number of infected agents and the county model is also able to capture a similar spread of the outbreak through the county. Further based on the analysis a switch value between 5% and 20% is likely going to produce the best results. A switch closer to 20% will better match the agent-based model but a switch closer to 5% will result in greater time savings and more time steps with an equation based disease component.

Further work can be done to improve the model, to make the model more realistic it should be tested where there is already some level of immunity in the population. This should require lower levels of switching and may produce different comparative results than what we have presented here. The difference equation model that we use for the equation based portion of the model is simple. There

are ways to create a more realistic equation based model, for example, adding additional equations for age groups. However, every additional equation makes the model more complicated and will include additional run time. The idea of creating a hybrid agent-based and equation based model is to simplify the model and save time and computing. Therefore, any work to further complicate the equation based model should keep that in mind.

Chapter 11

Testing an Intervention Strategy with the Hybrid Model

Chapter 10 presents a hybrid agent-based and equation based model and shows that such a model can be created so that its results do not significantly vary from a fully agent-based model. This allows us to reduce the fidelity of the model resulting in a model that requires less computing power and faster run times while not sacrificing results. Although this is an important step in the modelling process, it is also important to show that our model has the capability of producing results that can be useful. General agent-based model research can advance the field, however, creating a model with a specific disease, specific society that includes both maps and transportation allows us to do more than just agent-based model research and focus on disease dynamics research and epidemic planning. In Chapter 9, we use our county agent-based model to look at disease dynamics, specifically

how the centrality of a town in a network influences how an outbreak spreads to and from that town. In this chapter, we propose using our hybrid model and the results found in Chapter 9 to test intervention strategies focusing on school closure policies. We start the chapter with a brief overview of models looking at intervention strategies focusing on school closures, then discuss the model, our experiments and finally our results.

11.1 Agent-Based Models for Infectious Disease Interventions

One focus of many modelling studies is how interventions would influence the course of an outbreak. Modelling allows epidemiologists to test different interventions without a real world outbreak. The EpiSimdemics model is able to simulate detailed information on a disease spreading through a population including the individuals infected, where they were infected and who infected them. The information EpiSimdemics provides allows for identification of the severity of the epidemic as a whole and in certain subpopulations. The model has been used for multiple studies including those on pandemic planning for the US Department of Defence and the US Department of Health and Human Services. Looking at the effects of sequestering military sub-populations during a pandemic, the EpiSimdemics model determined that counter-intuitively sequestration may lead to more infections. It was determined this was because certain diseases can be infectious

before being symptomatic and although overall contacts would decrease with sequestration contacts in a smaller group of individuals, those who were sharing military quarters, would increase: resulting in infectious individuals being in close contact with susceptible individuals for a long period of time (Barrett et al., 2008).

An epidemiological intervention strategy that is occasionally implemented but its effectiveness is still debated is school closure policies. The policy of closing schools during disease outbreaks is used in Japan, Bulgaria and Russia to lessen influenza outbreaks (Litvinova et al., 2019). Its usefulness was debated in New York during the 1918 Spanish flu pandemic (Spinney, 2017) and is being used today to attempt to reduce the severity of measles outbreaks in Samoa (Kelly, November 23, 2019). However, there is no clear evidence to show that closing schools helps to reduce the size of an outbreak. In fact, Lee et al. (2008) find that shorter school closures of two weeks or less end up increasing the overall attack rate[1] and school closures may only be effective if they last for the entire duration of the epidemic. Similarly, Grefenstette et al. (2013) find that while the epidemic temporarily slows when schools close as soon as they reopen the epidemic peaks again. However, there are cases where studies have shown that school closure policies can play a significant role in reducing an outbreak. Litvinova et al. (2019) use real contact data to simulate the effects of a Russian school closure policy and find that reactive strategies, closing down classes and schools when a given percent of students show symptoms reduces the severity of an outbreak. We aim to use a model to test school closure policies that take into account a towns place in a

[1]Attack rate is defined as the risk of getting a disease during a specified period.

network of other towns: focusing on the centrality of other towns in the network and the physical distance to other towns.

11.2 Model

To test our intervention strategies we use the hybrid model presented in Chapter 10. The model is unchanged except to test intervention strategies we need to have the most realistic society that we can so we include vaccination. Thus our society component uses vaccination rates in the same way that is done in Chapter 9. Besides this the other components are identical to the hybrid model. As the introduction of vaccinations greatly reduces the total number of agents infected we use a smaller threshold when switching than used in the previous chapter. For the purposes of testing our interventions we use a threshold of 1% agents infected or exposed.

11.3 Experiments

Having a model that can recreate an outbreak allows us to learn interesting things about the dynamics of an infectious disease outbreak, for example in Chapter 9 we look into how the centrality of a town within a network of towns influences the spread of an outbreak and find that the agents commuting into a town are more important in spreading an infectious disease than the agents commuting out of a town and that the higher the in degree centrality of the town the outbreak

starts in the less important the centrality of the other towns in the network is in determining if the outbreak will spread to those other towns. However, even though this is an interesting finding, the question remains how can this help us in stopping or slowing down an outbreak. We propose using the findings to test out different intervention strategies. For example, while it might seem to make sense to close the schools in a town when an outbreak of a childhood disease begins to take off, there is evidence to show that this does not always help and in some cases actually makes an outbreak worse. However, as it was determined that a town with higher in degree centrality will result in greater spreading of the outbreak across all towns in the network, we run experiments to look at the effects of closing down schools in the high in degree centrality towns as opposed to the town the outbreak starts in. The thought behind this is that it is the high degree centrality towns that results in spreading to more towns throughout the network and that by stopping agents from going into these high centrality towns we will stop them from bringing the disease into the high centrality town and then out to other towns.

In order to test the effects of closing schools in different towns we run two experiments using the hybrid model. The first experiment involves four different interventions. The first is with no interventions but vaccination rates based off of Irish vaccination rates and with the outbreak starting in the town of Drumkeeran, Ireland in County Leitrim. Drumkeeran was chosen as it is a smaller town in Leitrim County with relatively low in degree centrality and it has both a primary and a secondary school in the town. The second set of runs uses the hybrid model

with vaccination rates and with schools in Drumkeeran closing down when more than two students are infected in the town. The third scenario again uses vaccinations and when more than two students are infected in Drumkeeran, the schools in Manorhamilton are shut down. Manorhamilton is the town in Leitrim with the highest degree centrality and is approximately 20 km away from Drumkeeran. The final set of initial conditions involves closing schools down in both Drumkeeran and Manorhamilton. The model is run for each set of initial conditions 300 times to account for stochasticity and we look at a number of different measures to compare the outbreaks.

The second experiment involves looking at different combinations of school closures based on the centrality of the towns versus the distance of the town to the initial outbreak. The outbreak begins in Drumkeeran for each set of runs. Looking at centrality, the initial set of runs closes down only the schools in Drumkeeran. Then in another set of runs we close schools in Drumkeeran and the town with the highest centrality that has at least one primary or secondary school. The next set of runs closes down schools in Drumkeeran, the town with the highest centrality, and the town with the second highest centrality. We continue with more runs adding school closures in towns with lower centrality. Table 11.1 shows the five towns in Leitrim containing at least one school with the highest in degree centralities in the county.

The second experiment looks at closing schools based on distance. This is done to help determine if any results from closing down schools by centrality is simply

Town	Centrality
Manorhamilton	197.8
Mohill	171.7
Drumshanbo	151.8
Carrigallen	93.7
Ballinamore	90.8

Table 11.1: Towns in Leitrim with at least one school ordered by centrality

because additional schools are closed or if it can be attributed to the centrality of the town that is closed. The first run closes schools in only Drumkeeran, the town where the outbreak starts. The next set of runs closes schools in Drumkeeran and the next closest town with at least one school in it. The third set of runs closes schools in Drumkeeran, the closest town with at least one school in it, and the second closest town with at least one school. Table 11.2 shows a list of the five closest towns to Drumkeeran.

Town	Distance (km)
Drumahaire	14
Killanummery	14
Drumshanbo	18
Manorhamillton	20
Leitrim	20

Table 11.2: Towns in Leitrim with at least one school ordered by distance to Drumkeeran

For each set of towns the model is run 300 times and the results are compared between closing schools based on distance and centrality.

11.4 Results

We first look at the results for the first experiment: vaccination, closing schools in Drumkeeran, closing schools in Manorhamilton and closing schools in Drumkeeran

and Manorhamilton. The first measure we look at to compare the results of the different interventions is to look at the percent of runs that lead to more than three agents infected. In a model without interventions we typically look at the percent of runs that lead to an outbreak, using the World Health Organization's (WHO) definition of a measles outbreak which is two or more connected cases of measles, however, because we want to look at the effects of the interventions, and the interventions do not start until we have at least two agents infected, we look at the runs where there are more than three agents infected. Table 11.3 shows the percent of runs that have three or more infected agents for each of the versions of the model along with the confidence intervals for the statistics.

Intervention	Percent of Runs	Confidence Interval
Vaccinations Only	51.3	(45.7, 57.0)
Schools Closed Drumkeeran	51.3	(45.7, 57.0)
Schools Closed Manorhamilton	47.3	(41.7, 53.0)
Schools Closed Drumkeeran and Manorhamilton	43.3	(37.8, 48.9)

Table 11.3: The percent of runs that result in three or more agents becoming infected based off of the intervention strategies used in the model

From Table 11.3 it can be seen that the percent of runs that lead to three or more infected agents is the same for the model with only vaccination as an intervention and the model closing schools in Drumkeeran. This further emphasises the findings showing that closing schools in the town where the outbreak starts does not always reduce an outbreak. Looking at the other interventions model we can see that even though the percent of runs with over three infected when schools are closed in Manorhamilton is slightly lower than the percent of runs when the schools are closed in Drumkeeran the results are not significantly different, with

each statistic in the other confidence interval. However, when schools close in both Drumkeeran and Manorhamilton, the percent of runs with more than three infected agents is significantly different than when the schools are closed in Drumkeeran. These results show that there might be an advantage in closing down the schools in the high centrality towns nearby along with the initial town where the outbreak starts.

The results are further broken down to see if there are other effects of closing schools. We look at some summary statistics for the county wide outbreak for all versions of the model in Table 11.4.

Intervention	1st Quartile	Median	Mean	3rd Quartile	Maximum
Vaccinations Only	1	4	27.01 *(23.95, 30.07)*	44	492
Schools Closed Drumkeeran	1	4	43.47 *(38.6, 48.4)*	45	4253
Schools Closed Manorhamilton	1	3	25.57 *(22.7, 28.5)*	39	652
Schools Closed Drumkeeran and Manorhamilton	1	2	23.62 *(21.0, 26.3)*	37.25	210

Table 11.4: Summary statistics for the models with interventions, including the confidence interval for the mean

From Table 11.4 we can see that there are some distinct differences between the results for the models. In particular when looking at the mean value across the runs the mean number of infected agents is significantly lower when schools are closed in Manorhamilton and when schools are closed in both Drumkeeran and Manorhamilton. In addition, the maximum value for the total number of infected agents across the 300 runs is much higher for when the model runs with schools closing in Drumkeeran versus when schools close in Manorhamilton or both Drumkeeran and Manorhamilton. Again we see this as a sign that closing the schools in the highly central towns may be a better option than closing the

schools in the town where the infection starts.

Additionally we look at how the outbreak spreads beyond the initial town within the network. The first measure that we look at is the number of cases of the disease outside of the initial town. Table 11.5 shows the percent of runs that lead to an outbreak (two or more infected) anywhere in the model and the percent of runs that have at least one agent infected outside of Drumkeeran. From the results and the confidence intervals we can see that there is no statistical difference between the runs when there is at least one agent infected outside of Drumkeeran for the model with only vaccination, the model where schools are closed in Drumkeeran or the model where schools are closed in Manorhamilton. However, we do see a statistically smaller result for the model where schools are closed in both Drumkeeran and Manorhamilton.

Intervention	Percent of Runs Outbreak	At Least one Infected Outside Drumkeeran
Vaccinations Only	63.3	49.7
	(57.9, 68.8)	*(44.0, 55.3)*
Schools Closed Drumkeeran	62.0	50.3
	(56.5, 67.5)	*(44.7, 56.0)*
Schools Closed Manorhamilton	61.7	47.3
	(56.2, 67.2)	*(41.7, 53.0)*
Schools Closed Drumkeeran and Manorhamilton	57.0	42.7
	(51.4,62.6)	*(37.1, 48.3)*

Table 11.5: A comparison between the percent of runs that lead to an outbreak (2 or more infected agents) and the percent of runs where at least one agent is infected from outside of Drumkeeran the initial town.

In the second set of experiments we look at closing schools based on their centrality. Schools initially close down in the town that the outbreak starts, Drumkeeran, then Drumkeeran and the town with the next highest centrality, Manorhamilton, then Drumkeeran, Manorhamilton and the town with the next

highest centrality, Mohill. We look at the percent of runs where three or more agents are infected. The results for this experiment are found in Table 11.6. From the table we can see that when we close down two schools there is a drop in the percent of runs that lead to an outbreak but after the schools in two towns have closed the percent rises again. This is an interesting finding but may be due to the fact that the students whose schools are closed do not change their behaviour in response to the outbreak. Instead of going to school they will treat the days off as if it were a weekend and thus will interact with each other potentially spreading the disease if an infected student decides to leave their home. Closing schools in three towns may be the tipping point from which closing schools reduces the outbreak to closing schools does not have an effect.

Additional Towns Closed	Percent of Runs	Confidence Interval
Drumkeeran	51.3	(45.7, 57.0)
Manorhamilton	43.3	(37.8, 48.9)
Mohill	47.3	(41.7, 53.0)
Drumshanbo	51.3	(45.7, 57.0)
Carrigallen	49.0	(43.3, 54.7)
Ballinamore	54.0	(48.4, 59.6)

Table 11.6: The percent of runs that result in three or more agents becoming infected when schools are closed by centrality.

To determine if the results we found have to do with the centrality of the town and not just the number of towns that the schools are closed in we also look at closing schools progressively by distance to the town where the outbreak starts. Table 11.7 shows the results for the percent of runs that lead to three or more infections closing towns by distance. The results show that similar to when we close schools by centrality, there is an initial decrease in the percent of runs that

have at least three agents infected. However, the drop in the percent is not as large as it is when the second town closed is based on centrality. This leads us to the conclusion that closing a second town might have a beneficial affect on reducing an outbreak but that closing schools in a town with high centrality is more beneficial than closing schools in a town with the closest distance.

Additional Towns Closed	Percent of Runs	Confidence Interval
Drumkeeran	51.3	(45.7, 57.0)
Drumahaire	45.3	(39.7, 51.0)
Killanummery	46.3	(40.7, 52.0)
Drumshanbo	49.3	(43.7, 55.0)
Manorhamilton	49.0	(43.3, 54.7)
Leitrim	46.0	(40.4, 51.6)

Table 11.7: The percent of runs that result in three or more agents becoming infected when schools are closed in towns by distance

11.5 Conclusion

We have used a hybrid agent-based and equation based model to look at the effects of different interventions on the outbreaks created in our model. It is important to be able to show that our hybrid model can not just produce model results but can also show us how an outbreak will change when different intervention strategies are used. Because testing intervention strategies is where we will be able to learn the most about an outbreak and how to prevent and slow one that occurs.

Although the literature is undecided about the usefulness of school closure policies on lessening the severity of an outbreak, it is still a commonly used strategy. We aimed to test interventions that look into stopping the outbreak from spreading out of the town of the initial case by looking at the schools in towns that have high

levels of in degree centrality, and schools that are close in distance to the initial town and found that we were able to reduce the severity of the outbreak spreading from Drumkeeran when we closed schools not only in the town where the outbreak begins but also in the town with the highest in degree centrality and the towns closest to Drumkeeran. From our finding we determined that closing schools in the town where the outbreak begins and then closing a second town is better at reducing the outbreak than just closing schools in the town where the outbreak initially occurs. In addition, we find that in selecting the second town in which to close schools, closing schools in a town with higher in degree centrality is more beneficial than closing schools in a town with a close distance to the source of the outbreak. This is likely because the high in degree centrality makes it more likely for an agent to commute to a town with higher centrality bringing the disease into the town. With the schools closed other agents will not become infected at school and bring the disease back to their own town thus reducing the spread. However, as we close schools in more towns both based on centrality and distance we see that with additional towns there is less of a reduction in outbreaks. This is likely because of an effect that is often cited as reason to not use a school closure policy, that the uninfected and asymptomatic students will still interact just outside of school and will still spread the disease.

Our findings show that closing down the schools in the town where an outbreak begins might not have as much of an effect on reducing the outbreak unless schools in another town are also closed: in particular closing the schools in towns with

the highest in degree centrality will result in the greatest decrease in the potential outbreaks.

These findings are the first step in coming up with intervention strategies to reduce outbreaks based off of town centrality. As we find that there is a reduction in potential outbreaks when a second town is closed regardless of if we chose the most central town or the closest town, further work could focus on looking at closing a combination of schools based on both centrality and distance. Additional work can be done to look at the results for different counties. Showing that our results work for Leitrim is one thing but running the same tests for other counties in Ireland or regions in other countries will show that the findings are robust and could be applied anywhere. We could also look at different thresholds for closing down schools: instead of closing down when two agents are infected we could wait for a larger number of students to be infected before the schools close down to determine if this threshold has an impact on the results and which intervention strategies work the best. There is also the potential to look into changing the agents behaviours after schools close. Instead of moving as if it is the weekend agents could adjust their actions to prevent transmission knowing that there is an outbreak occurring. Similarly, in the current version of the model only students actions are changed when schools close down but the actions of adults whose children attend schools that closed down could also be adjusted. With such adjustment it might also be possible to calculate the economic impact of closing schools down versus letting an outbreak run its course without interventions. This could include the cost of

paying teachers salaries while the school is closed and the cost for parents taking the days off from work compared to the cost of treating the number of agents who would be infected. Although closing down schools in two towns seems to have a beneficial affect on reducing an outbreak, in order to adopt such a policy it would need to be shown that the reduction in the outbreak was not outweighed by the cost of closing down the schools.

Chapter 12

Conclusion

This chapter provides a summary of the work presented in this thesis (Section 12.1), then discusses the main research questions addressed and how they were answered (Section 12.2), before finally discussing directions for future work (Section 12.3).

12.1 Summary

Agent-based modelling is a "bottom up" modelling method that can simulate a real world system in high level of detail. This detail allows for a greater understanding of the world around us through analysing the results of the model. This is particularly important for modelling the spread of infectious diseases as it is individual actions and interactions that are important in determining the course of an outbreak especially at the early stages. However, a highly detailed and specific model can take a lot of computing power and thus time to run. This time constraint could be a deterrent in running an agent-based model especially in a

situation that requires quick action such as in the case of an outbreak that has already begun. For these reasons, equation based models which have been shown to capture macro level disease dynamics and are much more time efficient are often chosen. However, equation based models do not capture the individual actions and differences that make the use of an agent-based model attractive.

In this work we have proposed using a hybrid agent-based and equation based model to reduce the time needed to run an agent-based model while still capturing the individual level detail that is not seen in an equation based model. Our model discussed in Chapter 10 is created with a detailed specific society component, a specific disease component, and includes transportation and maps in the environment. As such, within the taxonomy of agent based models presented in Chapter 4 our model is a specific model, and is therefore suitable for epidemiology planning. Indeed, in Chapter 11 we use the model to investigate how school closure policies can influence an outbreak. In our work we first analysed both equation based models (Chapter 3) and agent-based models (Chapter 5) for the spread of infectious diseases within the context of a small town. While both models are easily adaptable to model different towns, the agent-based model has some advantages. It is more easily adaptable to intervention strategies in particular to those strategies that involve changes in agents' behaviours during the course of the outbreak. Agent-based models also allow for individual characteristics to be given to each agent whereas in an equation based model all agents within the same compartment are homogeneous.

It is also important to note that the stochasticity in an agent-based model which comes from the agents making their own decisions produces different results each model run. Therefore, if the model is run many times a range of possible outbreaks is determined. This range of outbreaks is not found in an equation based model as they typically lack stochasticity unless stochasticity is added in as an additional element. However, the range of outbreaks is important in understanding what might actually occur in a real outbreak. Agent-based models are able to give us different possible scenarios for what might happen and can show how an intervention, such as a change in vaccination rates might reduce the size of an outbreak but also the chance that an outbreak will occur. An equation based model is only able to give us the change in size. Because of these advantages we decided to focus our main work on agent-based models.

Focusing on agent-based models we built a detailed agent-based model using only openly available data in order to test the ability of an agent-based model to recreate a highly detailed system and to determine how and if the greater level of detail will have an effect on the model results (Chapter 5). With the model in Chapter 8 we were able to show that clustering of individuals by socioeconomic status can result in larger outbreaks and a higher percent chance that an outbreak will occur. However, even though the detailed model shows that there can be advantages in including more data and more detail in our model the trade off of detail versus computing power needs to be considered. We could continue to add more and more data and detail to our model but at a point it would become only

feasible to model a small population because the detail is so high.

While there is some information that we can learn from modelling smaller populations, ultimately it is important to look at commuting and travel between those smaller populations or towns as that travel will play an important role on the spread of an outbreak and no town exists in isolation. In addition, although increasing the complexity can result in differences in results, such as the case with the burn-in model from Chapter 8, there is not a large change in the results and there is a large increase in run time. This increase in run time combined with only a small increase in results and the inability of a closed town model to capture the commuting and travel that can be important to the spread of an infectious disease led us to move towards a regional model that used a number of assumptions to scale up from the single town model. Although we show that our trade-offs do not drastically change the results there are still some trade-offs that need to be considered when scaling up from a single town to a network of towns: should fidelity be persevered over run time or where does run time become more important than fidelity? The work on scaling up the agent-based model in Chapter 9 focuses on this question while looking at the additional information we can learn about a towns susceptibility to an outbreak with a scaled up model.

We further examine the trade-off of balancing fidelity with saving computing power and time when we create a hybrid agent-based model and equation based model to replace our fully agent-based model described in Chapter 10. The hybrid model switches in the disease component of the model only. This decision allows

us to keep the movement of the agents as agent-based and allows for agents to retain their own individual characteristics. The decisions allow us to retain many of the advantages of the agent-based model while saving time by using the equation based model.

In order to properly test the hybrid we looked at different geographic levels for switching and at different thresholds for switching. This analysis allowed us to fully understand how the results of our model changed as we switch the disease component between agent-based and equation based over larger or smaller geographies (county, town, or small area) and at different thresholds, requiring different numbers of agents to be infected before switching. From this, as expected, we find that a higher threshold, meaning more agents have to be infected to switch to an equation based disease component, results in the model staying fully agent-based for a longer period of time and preserves more fidelity. We also find that the fidelity of the model is better preserved when the disease component switches by town.

These trade-offs of fidelity versus computing time are important to consider when creating a model. While a detailed model can provide accurate results by not leaving out important factors that can influence the spread of a disease, a very detailed model can take an inordinate amount of time to run and a model that needs weeks to run might not be useful especially in a situation with time constraints. Thus a trade-off with fidelity must be considered and assumptions about what factors are the most important must be made. For example, in order

to scale up to a county level agent-based model we assumed that the network of towns an agent can move between is more important to consider and provides more information on the spread of a disease than fine level socioeconomic segregation. Similarly, to save additional time and computing power we decided that the individual movements and decisions that influence the disease component of the model are most important when a small number of agents are infected and are less important when a large number of agents are infected so switching from an agent-based to an equation based disease component does not result in a large loss of fidelity. And we are able to create a model that can run in less time and still be confident in the results. We are even still able to use the model to show how intervention strategies can help to mitigate an outbreak.

Another large part of the thesis that was done concurrently with creating our model was creating a methodology for implementing and evaluating an agent-based model. We propose to use the taxonomy created from an in depth analysis of the literature and presented in Chapter 4 to aide in the creation and implementation of new agent-based models. In addition, Chapter 7 gives the steps we took to evaluate our model and generalizes them so that they can be applied to other agent-based models.

12.2 Research Questions

Taken together these research contributions address the two main research questions proposed in Chapter 1. Our first research question asks how we can create

a model of the spread of disease in a specific society that is able to capture the behaviour and interactions of heterogeneous agents while taking into consideration the computing power needed to run the model. This thesis outlines the steps taken to create such a model from comparing basic equation based and agent-based models to scaling up the model and eventually creating a hybrid model. Although research often focuses on agent-based or equation based models and the advantages and disadvantages of each more work should be done on ways to combine the advantages of both. Hybrid models are a way to preserve the heterogeneous movements and characteristics of agents while simplifying the model. However, in order to decide how to use a hybrid model one needs to understand the system being modelled and the performance of both models. For example, in order to not lose too much fidelity the hybrid model presented in this thesis switches between agent-based and equation based for only the disease component of the model. This allows the model to still capture the heterogeneous movements of agents that we found was important when looking at the centrality of the towns in the county model. The hybrid model allows for the centrality of the towns to remain accurate based off of the commuting patterns of agents. Without going through the process of creating an agent-based model on a smaller scale, and then scaling up that agent-based model, it would be more difficult to determine what parts of the model should remain agent-based and what parts could switch to equation based. Thus while a hybrid model can be useful and save computing power while retaining some heterogeneous characteristics of agents and still produce emergent results, it

is essential to go through the process of creating a validated agent-based model to first understand the system.

In addition, to further answer the first research questions we focus on creating a standard and methodology for agent-based models for infectious disease epidemiology. The flexibility of an agent-based model is a strong advantage in choosing the method but it can also be a disadvantage with no clear methodology for creating and evaluating models. Our taxonomy presented in Chapter 4 aims to fill the gap in helping to classify and guide creation of models while the methodology for validating models presented in Chapter 7 can help to show that a model has been properly validated and can be used to make conclusions about the system. Without a defined methodology, it is difficult to show that a model has been properly evaluated. It is important for modellers to take the steps to determine what their methodology is and useful to have a general methodology to follow as a guide.

The second research question involves applying the model to plan for future outbreaks. As we created a model that can be used for outbreak planning based on our taxonomy it should have a specific society, and specific disease. This guided our model creation but also makes it essential to determine if we can use the model to learn about outbreaks. This was done throughout model development looking at how the models respond to changes in vaccination policy that alter the initial conditions of the model in Chapter 6, looking into how socioeconomic clustering can influence an outbreak in Chapter 8, determining how the centrality of a town influences the spread of the outbreak in Chapter 9 and looking at school

closure policies in Chapter 11. Planning for an outbreak can involve more than testing out potential intervention strategies such as school closure policies, it is also important to learn as much as possible about the factors involved in a potential outbreak before one occurs. Many times it is not one characteristic of a town, county or population that makes it more or less susceptible to an outbreak but a combination of characteristics. Our model can help to determine what those combinations are. Although creating a validated model is important and can show that we should be able to apply the results of the model to a real world system, actually applying model results to the real world and making conclusions about those results is essential in showing that a model is valuable.

While answering our research questions in this thesis we made the following contributions to the state of the art:

- A taxonomy of agent-based models for human infectious disease epidemiology. The taxonomy is presented in Chapter 4 and published in Hunter et al. (2017).

- A methodology for validating and testing an agent-based model for the spread of infectious diseases is presented in Chapter 7.

- A burn-in segregation model as an extra step in the agent-based model setup is presented in Chapter 8 and published in Hunter et al. (2018c).

- A hybrid agent-based and equation based model architecture that combines the two existing hybrid modelling methods that is discussed in Chapter 10.

- An agent-based model for the Irish context for the spread of measles: the fully agent-based model for Irish towns is found in Chapter 5 and published in Hunter et al. (2018b), the fully agent-based model for Irish counties is found in Chapter 9 and the hybrid agent-based and equation based model for Irish counties can be found in Chapter 10.

- An analysis of how the centrality of a town within a network of other towns affects the spread of an infectious disease through the network in Chapter 9 and published in Hunter et al. (2019).

- An analysis of different school closure policies using the results from our centrality analysis in Chapter 11.

12.3 Future Work

At this point the model has only been run for a county in Ireland. Future work can involve extending the model to other counties in Ireland and counties and regions in other countries. This could allow us to determine if the results that we found for Leitrim, Ireland are robust across other counties and regions. Additionally, we could scale up the model further to model an entire country. Along with scaling up the model we can look at some additional intervention measures. In our work we have looked at how different levels of vaccinations affect an outbreak and how school closure policies might affect an outbreak. However, we could look at further options such as reactive vaccination campaigns where individuals who are

not vaccinated are targeted once an outbreak reaches a certain size or the closure of public spaces and events.

The focus of our model in this work is on measles, however the model could be easily extended to model other person to person and airborne infectious diseases such as influenza or mumps. It could be important to understand how outbreaks vary based on different infectious diseases or how certain population characteristics might make a population more or less susceptible to a given disease. Beyond diseases with the same infection mechanisms pieces of our model could be used to study other types of infectious or non infectious diseases. If we adjust the disease component and the environmental component the model could be used to simulate the spread of a water borne infectious disease within a population and using the society component as a base we could add some additional characteristics and change the time scale of the model to simulate the development of non communicable diseases based off of individual and environmental characteristics.

References

Ajelli, M., Goncalves, B., Balcan, D., Colizza, V., Hu, H., Ramasco, J. J., ... Vespignani, A. (2010). Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. *BMC Infectious Diseases*, *10*(190). Retrieved from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3306662/` doi: 10.1186/1471-2334-10-190

Aleman, D. M., Wibisono, T. G., & Schwartz, B. (2011). A nonomogeneous agent-based simulation approach to modeling the spread of disease in a pandemic. *informs*, *41*(3), 301-315. Retrieved from `https://www.researchgate.net/publication/220249824` doi: 10.1287/inte.1100.0550

Allen, L. J. S. (1994). Some discrete-time si, sir, and sis epidemic models. *Mathematical Biosciences*.

Apolloni, A., Kumar, V. A., Marathe, M. V., & Swarup, S. (2009). Computational epidemiology in a connected world. *IEEE Computer Society*, 97–100. Retrieved from `http://staff.vbi.vt.edu/swarup/papers/computational-epidemiology.pdf` doi: 10.1109/MC.2009.386

Barrett, C. L., Bisset, K. R., Eubank, S. G., Feng, X., & Marathe, M. V. (2008). Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. *International Conference for High Performance Computing, Networking, Storage and Analysis*. Retrieved from http://dx.doi.org/10.1109/SC.2008.5214892 doi: 10.1109/SC.2008.5214892

Bartlett, P., & Judge, L. J. (1997). The role of epidemology in public health. *Scientific and Technical Review of the Office International des Epizooties*.

Benenson, I., Hatna, E., & Or, E. (2009). From schelling to spatially explicit modeling of urban ethnic and economic residential dynamics. *Sociologial Methods and Research*. doi: 10.1177/0049124109334792

Binder, B. J., Ross, J. V., & Simpson, M. J. (2012). A hybrid model for studying spatial aspects of infectious diseases. *The ANZIAM Journal*, *54*, 37-49. doi: https://doi.org/10.1017/S1446181112000296

Bobashev, G. V., Goedecke, D. M., Yu, F., & Epstein, J. M. (2007). A hybrid epidemic model: Combining the advantages of agent-based and equation based-approaches. *Proceedings of the 2007 Winter Simulation Conference*, 1532-1537. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4419767&tag=1 doi: 10.1109/WSC.2007.4419767

Bradhurst, R. A., Roche, S. E., East, I. J., Kwan, P., & Garner, M. G. (2015). A hybrid modeling approach to simulating foot-and-mouth disease outbreaks in

australian livestock. *Fronteirs in Environmental Science*. doi: 10.3389/fenvs .2015.00017

Bradhurst, R. A., Roche, S. E., Garner, M. G., Sajeev, A. S. M., & Kwan, P. (2013). Modelling the spread of livestock disease on a national scale: the case for a hybrid approach. *20th International Congress on Modelling and Simulation*.

Chakraborty, A., Wilson, K., Green, N., Alur, S. K., Ergin, F., Gurumurthy, K., ... Chinta, D. (2014). Practical graph mining with r. In (chap. Link Analysis). CRC Press.

Clark, W. A. V. (1991). Residential preferences and neighborhood racial segreation: A test of the schelling segregation model. *Demography*. doi:

Clark, W. A. V., & Fossett, M. (2008). Understanding the social context of the schelling segregation model. *PNAS*. doi: 10.1073/pnas.0708155105

Clarke, V. (2018). Mandatory measles vaccinations for children could lead to backlash, professor warns. *The Irish Times*.

Coffee, N., & Lockwood, T. (2012). The property wealth metric as a measure of socioeconomic status. *18th Annual PRRES Conference*. Retrieved from `http://www.prres.net/papers/coffee_property_wealth_metric_for_se_status.pdf`

Corner, S., Ryan, F., MacSweeney, M., Coughlan, H., & Kieran, M. (2012). Measles outbreak west cork may 2012. *Epi-Insight*. Re-

trieved from http://ndsc.newsweaver.ie/epiinsight/dllmrzt7dc0?a=1&p=24661425&t=17517774 doi:

Crooks, A. T., & Hailegiorgis, A. B. (2014). An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling & Software*, *62*, 164 - 177. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364815214002515 doi: 10.1016/j.envsoft.2014.08.027

CSO. (2014a). *Census 2011 boundary files.* Retrieved from http://www.cso.ie/en/census/census2011boundaryfiles/ (Date accessed 26.05.2016)

CSO. (2014b). Census 2011 boundary files. *Central Statistics Office*. Retrieved from http://www.cso.ie/en/census/census2011boundaryfiles/

CSO. (2017). Census 2016 place of work, school or college - census of anonymised records (powscar). *Central Statistics Office*. Retrieved from https://www.cso.ie/en/census/census2016reports/powscar/

Dibble, C., Wendel, S., & Carle, K. (2007). Simulating pandemic influenza risks of us cities. *Proceedings of the 2007 Winter Simulation Conference*, 1548-1550. Retrieved from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4419770&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4419770 doi: 10.1109/WSC.2007.4419770

DOE. (2017). Data on individual schools. *Department of Education of Education and Skills*. Retrieved from http://www.education.ie/en/Publications/

Statistics/Data-on-Individual-Schools/Data-on-Individual-Schools.html

Doherty, E., Walsh, B., & Neill, C. O. (2014). Decomposing socioeconomic inequality in child vaccination: Results from ireland. *Vaccine*. doi: 10.1016/j.vaccine.2014.03.084

Duan, W., Fan, Z., adn Gang Guo, P. Z., & Qiu, X. (2015). Mathematical and computational approaches to epidemic modeling: A comprehensive review. *Frontiers of Computer Science*, •. doi: 10.1007/s11704-014-3369-2

Duan, W., Qiu, X., Cao, Z., Zheng, X., & Cui, K. (2013). Heterogeneous and stochastic agent-based models for analyzing infectious diseases' super spreaders. *IEEE Intelligent Systems*, *13*, 1541-1672. Retrieved from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6468033 doi: 10.1109/MIS.2013.29

Dunham, J. B. (2005). An agent-based spatially explicit epidemiological model in mason. *Journal of Artificial Societies and Social Simulation*, *9*(1), 3.

Endrich, M. M., Blank, P. R., & Szucs, T. D. (2009). Influenza vaccination uptake and socioeconomic determinants in 11 european countries. *Vaccine*. doi: 10.1016/j.vaccine.2009.04.029

Epstein, J. M., Parker, J., Cummings, D., & Hammond, R. A. (2008). Coupled contagion dynamics of fear and disease: Mathematical and computational explorations. *PLOS one*, *3*(12), 1-11. Retrieved from http://journals.plos.org/

plosone/article?id=10.1371/journal.pone.0003955 doi: 10.1371/journal.pone.0003955

Fossett, M. (2011). Generative models of segregation: Investigating model-generated patterns of residential segregation by thnicity and socioeconomic status. *Journal of Mathematical Sociology*.

Friás-Martínez, E., Williamson, G., & Friás-Martínez, V. (2011). An agent-based model of epidemic spread using human mobility and social network information. *IEEE Conference on Social Computing*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6113095 doi: 10.1109/PASSAT/SocialCom.2011.142

Gilbert, N. (2008). *Agent-based models*. London: Sage Publications, Inc.

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., ... Burke, D. S. (2013, Oct 08). Fred (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health*, *13*(1), 940. Retrieved from https://doi.org/10.1186/1471-2458-13-940 doi: 10.1186/1471-2458-13-940

Grimm, V., Berger, U., Deangelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The odd protocol: A review and first update. *Ecological Modelling*. doi: 10.1016/j.ecolmodel.2010.08.019

Hegselmann, R. (2017). Thomas c. schelling and james m. sakoda: The intellectual, technical, and social history of a model. *Journal of Artificial Societies and Social Simulation*, *20*(3), 15. Retrieved from http://jasss.soc.surrey.ac.uk/20/3/15.html doi: 10.18564/jasss.3511

Hernán, M. A. (2014). Invited commentary: Agent-based models for casual inference –reweighting data and theory in epidemiology. *American Journal of Epidemiology*, *181*(2), 103–105. Retrieved from http://aje.oxfordjournals.org/content/181/2/103.full doi: 10.1093/aje/kwu272

Hethcote, H. W. (1989). Applied mathematical ecology. biomathematics. In (chap. Three Basic Epidemiological Models). Springer.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *Society for Industrial and Applied Mathematics Review*, 599 - 653.

Hogan, A. B., Glass, K., Moore, H. C., & Anderssen, R. S. (2016). Applications + practical conceptualization + mathematics = fruitful innovation:age structures in mathematical models for infectious diseases, with a case study of respiratory syncytial virus. *Mathematics for Industry*, 105-116. doi:

HPSC. (2017). *Immunisation uptake statistics at 12 and 24 months of age.* http://www.hpsc.ie/a-z/vaccinepreventable/vaccinationimmunisationuptakestatistics/immunisationuptakestatisticsat12and24monthsofage/.

HSE. (2017). *Health a-z:measles.* Retrieved from http://www.hse.ie/eng/health/az/M/Measles/

HSE. (June 2012). Measles outbreak west cork. *Immunisation Focus*.

HSE. (September 2012). Measles in west cork: Outbreak over. *Immunisation Focus*.

Hunter, E., Mac Namee, B., & Kelleher, J. (2018b, 12). An open-data-driven agent-based model to simulate infectious disease outbreaks. *PLOS ONE*, *13*(12), 1-35. Retrieved from https://doi.org/10.1371/journal.pone.0208775 doi: 10.1371/journal.pone.0208775

Hunter, E., Mac Namee, B., & Kelleher, J. D. (2017). A taxonomy for agent-based models in human infectious disease epidemiology. *Journal of Artificial Societies and Social Simulation*, *20*(3), 2. Retrieved from http://jasss.soc.surrey.ac.uk/20/3/2.html doi: 10.18564/jasss.3414

Hunter, E., Mac Namee, B., & Kelleher, J. D. (2018a). A comparison of agent-based models and equation based models for infectious disease epidemiology =. In *Proceedings of the 26th aiai irish conference on artifical intelligence and cognitive science.*

Hunter, E., Mac Namee, B., & Kelleher, J. D. (2018c). Using a socioeconomic segregation burn-in model to initialise an agent-based model for infectious diseases. *Journal of Artificial Societies and Social Simulation*, *21*(4), 9. Retrieved from http://jasss.soc.surrey.ac.uk/21/4/9.html doi: 10.18564/jasss.3870

Hunter, E., Namee, B. M., & Kelleher, J. D. (2019). *Degree centrality and the probability of an infectious disease outbreak in towns within a region.*

Iceland, J., Weinberg, D. H., & Steinmetz, E. (2002). Racial and ethnic residential sergregation in the united states: 1980-2000. *Census 2000 Special Reports: US Census Bureau*. Retrieved from `https://www.census.gov/prod/2002pubs/censr-3.pdf` doi:

IDMDocs. (2019). *Seir and seirs models.* Retrieved from `http://www.idmod.org/docs/general/model-seir.html`

Jessop, L. J., Murrin, C., Lotya, J., Clarke, A. T., O'Mahony, D., Fallon, U. B., ... Murphy, A. W. (2010). Socio-demographic and health related predictors of uptake of first mmr immunisation in the lifeways cohort study. *Vaccine*. doi: 10.1016/j.vaccine.2010.06.092

Kao, R. R. (2002). The role of mathematical modelling in the control of the 2001 fmd epidemic in the uk. *TRENDS in Microbiology*, 279-286. doi:

Kasereka, S., Kasoro, N., & Chokki, A. P. (2014, Nov). A hybrid model for modeling the spread of epidemics: Theory and simulation. In *2014 4th international symposium isko-maghreb: Concepts and tools for knowledge management (isko-maghreb)* (p. 1-7). doi: 10.1109/ISKO-Maghreb.2014.7033457

Keeling, M., & Grenfell, B. (2000). Individual-based perspectives on r_0. *Journal of Theoretical Biology*. doi: 10.1006/jtbi.1999.1064

Keeling, M. J., & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton: Princeton University Press.

Kelly, L. (November 23, 2019). Samoa measles outbreak kills 20, mostly children. *Reuters*.

Ketema, D., Mamo, & Koya, P. R. (2015). Mathematical modeling and simulation study of seir disease and data fitting of ebola epidemic spreading in west africa. *Journal of Multidisciplinary Engineering Science and Technolgy*, 106-114. doi:

Lee, B. Y., Brown, S. T., Cooley, P., Potter, M. A., Wheaton, W. D., Voorhees, R. E., ... Burke, D. S. (2008). Simulating school closure strategies to mitigate an influenza epidemic. *Journal of Public Health Managment and Practice*, *16*(3), 252-261. Retrieved from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2901099/` doi: 10.1097/PHH.0b013e3181ce594e

Linard, C., Poncon, N., Fontenille, D., & Lambin, E. F. (2009). A multi-agent simulation to assess the risk of malaria re-emergence in southern france. *Ecological Modelling*, *220*(2), 160 - 174. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0304380008004304` doi: 10.1016/j.ecolmodel.2008.09.001

Litvinova, M., Liu, Q.-H., Kulikov, E. S., & Ajelli, M. (2019). Reactive school closure weakens the network of social interactions and reduces the spread of influenza. *Proceedings of the National Academy of Sciences*, *116*(27), 13174–

13181. Retrieved from https://www.pnas.org/content/116/27/13174 doi: 10.1073/pnas.1821298116

Mackenbach, J. P., Meerding, W. J., & Kunst, A. E. (2007). *Economic implications of socio-economic inequalities in health in the european union* (Tech. Rep.). Department of Public Health, Rotterdam, The Netherlands: Health and Consumer Protection Directorate-General.

Mao, L. (2014). Modeling triple-diffusions of infectious diseases, information, and preventive behaviors through a metropolitan social network – an agent-based simulation. *Applied Geography*, *50*, 31 - 39. Retrieved from http://www.sciencedirect.com/science/article/pii/S0143622814000277 doi: 10.1016/j.apgeog.2014.02.005

Marilleau, N., Lang, C., & Giraudoux, P. (2018). Coupling agent-based with equation-based models to study spatilly explicit megapopulation dynamics. *Ecological Modelling*. doi: https://doi.org/10.1016/j.ecolmodel.2018.06.011

Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social Forces*. doi: 10.2307/2579183

MATLAB. (2017). *version 9.2.0 .556344(r2017a).* Natick, Massachusetts: The MathWorks Inc.

Measeles: Vaccine-preventable diseases surveillance standards (Tech. Rep.). (2018). World Health Organization.

Merler, S., Ajelli, M., Fumanelli, L., Gomes, M. F. C., y Piontti, A. P., Rossi, L., ... Vespignani, A. (2015). Spatiotemporal spread of the 2014 outbreak of ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*, *15*(2), 204 - 211. Retrieved from http://www.sciencedirect.com/science/article/pii/S1473309914710746 doi: 10.1016/S1473-3099(14)71074-6

Moudon, A., Cook, A., Ulmer, J., Hurvitz, P., & Drewnowski, A. (2011). A neighborhood wealth metric for use in health studies. *American Journal of Preventative Medicine*. doi: 10.1016/j.amepre.2011.03.009

Muldoon, R., Smith, T., & Weisberg, M. (2012). Segregation that no one seeks. *Philosophy of Science*. doi: 10.1086/663236

Myplan.ie. (2017). Zoning data. *Department of Housing, Planning, Community and Local Government*. Retrieved from http://myplan.ie/index.html

National security capability review (Tech. Rep.). (2018). Cabinet Office. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/705347/6.4391_CO_National-Security-Review_web.pdf

Nelson, K. E., & Williams, C. M. (2007). *Infectious disease epidemiology theory and practice*. Jones and Bartlett Publishers.

Newstalk. (2017). Kinsale named town that is most representative of the nation.

•. Retrieved from https://www.newstalk.com/Kinsale-named-town-that-is-most-representative-of-the-nation

Nicoara, C., Zach, K., Trachsel, D., Germann, D., & Matter, L. (1999). Decay of passively acquired maternal antibodies against measles, mumps and rubella viruses. *Clinical and Vaccine Immunology*. doi:

O'Brien, S., & Cotter, S. (2018). Measles: an ongoing threat to public health in ireland and europe. *Epi-Insight: Disease Surveillance Report of HPSC, Ireland*.

O'Connor, B., Cotter, S., Heslin, J., Lynam, B., McGovern, E., Murray, H., ... Doyle, S. (2016). Catching measles in an appopriately vaccinted group: A well-circumscribed outbreak in the south east of ireland, september-november 2013. *Epidemiology and Infection*. Retrieved from doi: 10.1017/S095026881600145X

OECD. (2017). *Child vaccination rates (indicator).* doi: 10.1787/b23c7d13-en

Olsen, J., & Jepsen, M. R. (2010). Human papillomavirus transmission and cost-effectiveness of introducing quadrivalent hpv vaccination in denmark. *International Journal of Technology Assesment in Health Care*, *26*(2), 183 – 191. Retrieved from http://dx.doi.org/10.1017/S0266462310000085 doi: 10.1017/S0266462310000085

OpenStreetMap contributors. (2017). *Planet dump retrieved from https://planet.osm.org* . https://www.openstreetmap.org.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted

networks: Generalizing degree and shortest paths. *Social Networks*, *245-251*(3), 1-35. Retrieved from https://doi.org/10.1016/j.socnet.2010.03.006

Pang, L., Ruan, S., Liu, S., Zhao, Z., & Zhang, X. (2014). Transmission dynamics and optimal control of measles epidemics. *Applied Mathematics and Computation*, 131-147. doi:

Perez, L., & Dragicevic, S. (2009). An agent-based approach for modeling dynamics of contagious disease spread. *International Journal of Health Geographics*. doi: 10.1186/1476-072X-8-50

PSRA. (2012). Residential property price register. *Property Services Regulatory Authority*. Retrieved from http://www.psr.ie/website/npsra/npsraweb.nsf/page/index-en

QGIS. (2009). Qgis geographic information system 2.8. *Open Source Geospatial Foundation*. Retrieved from http://www.qgis.org/en/site/index.html

Rakowski, F., Gruziel, M., Bieniasz−Krzywiec, L., & Radomski, J. P. (2010a). Influenza epidemic spread simulation for poland — a large scale, individual based model study. *Physica A: Statistical Mechanics and its Applications*, *389*(16), 3149 - 3165. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437110003687 doi: 10.1016/j.physa.2010.04.029

Ratner, B. (2009). The correlation coefficient: Its values range between +1/1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing*. doi:

Richiardi, M., Leombruni, R., Saam, N. J., & Sonnessa, M. (2006). A common protocol for agent-based social simulation. *Journal of Artificial Societies and Social Simulation*, *9*(1), 15. Retrieved from `http://jasss.soc.surrey.ac.uk/9/1/15.html`

Ridenhour, B., Kowalik, J. M., & Shay, D. K. (2014). Unravelling r_0: Considerations for public health applications. *American Journal of Public Health*. doi: 10.2105/AJPH.2013.301704

Rodrigue, J.-P., Comtois, C., & Slack, B. (2006). *The geography of transport systems*. London: Routledge, Taylor and Francis Group.

Schelling, T. C. (1969). Models of segregation. *The American Economic Review*.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*.

Seither, R., Masalovich, S., Knighton, C. L., Mellerson, J., Singleton, J. A., & Greby, S. M. (2014). Vaccination coverage among children in kindergarten united states, 2013-2014 school year. *Morbidity and Mortality Weekly Report*.

Simoes, J. M. (2006). Modelling a mumps outbreak through spatially explicit agents. *Potentials of Complexity Science for Business, Governments, and the Media 2006*. Retrieved from `http://www.casa.ucl.ac.uk/joanamargarida/papers/SimoesJ.pdf`

Skvortsov, A. T., Connell, R. B., Dawson, P. D., & Gailis, R. M. (2007). Epi-

demic modelling: Validation of agent-based simulation by using simple mathematical models. *International Congress on Simulation and Modelling*, 657–662. Retrieved from http://mssanz.org.au.previewdns.com/MODSIM07/papers/13_s20/EpidemicModeling_s20_Skvortsov_.pdf

Spinney, L. (2017). *Pale rider: The spanish flu of 1918 and how it changed the world*. PublicAffairs.

Stoica, V. I., & Flache, A. (2014). From schelling to schools: A comparison of a model of residential segregation with a model of school segregation. *Journal of Artifical Societies and Social Simulation*.

Taubenberger, J. K., & Morens, D. M. (2006). 1918 influenza: the mother of all pandemics. *Emerging Infectious Diseases*. doi: 10.3201/eid1201.050979

Thomas, J. C., & Weber, D. J. (2001). Epidemiologic methods for the study of infectious diseases. In (p. 61-62). Oxford University Press.

Tian, Y., Osgood, N. D., Al−Azem, A., & Hoeppner, V. H. (2013). Evaluating the effectiveness of contact tracing on tuberculosis outcomes in saskatchean using individual-based modeling. *Health Education and Behavior*, *40*(15), 985-1105. Retrieved from http://heb.sagepub.com/content/40/1_suppl/98S.short doi: 10.1177/1090198113493910

Vaidya, N. K., Morgan, M., Jones, T., Miller, L., Lapin, S., & Schwartz, E. J. (2015). Modelling the epidemic spread of an h1n1 influenza outbreak in a rural university town. *Epidemiology and Infection*, 1610 - 1620. doi:

Wheaton, W. D., Cajka, J. C., Chasteen, B. M., Wagener, D. K., Cooley, P. C., laxminarayana Ganapathi, ... Allpress, J. L. (2009). Synthesized population databases: A us geospatial database for agent-based models. *RTI International*. Retrieved from doi: 10.3768/rtipress.2009.mr.0010.0905

WHO. (2018). 2018 annual review of diseases prioritized under the research and development blueprint. *WHO Research and Development Blueprint*.

Wilensky, U. (1999a). Netlogo. *Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.*. Retrieved from `https://ccl.northwestern.edu/netlogo/`

Wilensky, U. (1999b). Netlogo segregation model. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.*, •.

Xia, H., Barrett, C., Chen, J., & Marathe, M. V. (2013). Computational methods for testing adequacy and quality of massive synthetic proximity social networks. *CSE '13 Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering*, 1113-1120. Retrieved from `http://staff.vbi.vt.edu/chenj/pub/NDSSL-TR-13-153.pdf`

Yoneyama, T., Das, S., & Krishnamoorthy, M. (2012). A hybrid model for disease spread and an application to the sars pandemic. *Journal of Artificial Societies and Social Simulation*, *15*(1), 5. Retrieved from `http://jasss.soc.surrey.ac.uk/15/1/5.html` doi: 10.18564/jasss.1782

Appendices

Appendix A

Town Model Description

Model Description

The model presented in this description is a model for measles in an Irish town, it is presented in Chapters 5 and 8. We use the computer software Netlogo (Wilensky, 1999a) to implement our model. Netlogo is an easy to use and popular environment for creating agent-based models (Gilbert, 2008). We chose to use Netlogo due to the increasing popularity of the platform with agent-based modellers and its ability as a medium to high/large scale modelling platform. It does, however, have disadvantages, one of the biggest being the speed of the program. When modelling simulations with a small number of agents Netlogo works well, however, once the number of agents gets large enough the simulation slows down. In our current model we are able to easily simulate towns of up to 10,000 agents. However, beyond that the model begins to break down. The open-data-driven approach described in this paper, however, could be used with any agent-based modelling

tool or platform.

Our model is a data driven agent-based model for human airborne infectious diseases such as the flu or measles. It follows an SEIR (susceptible, exposed, infected, and recovered) type compartmental model with the agents moving between the four state relating to infectiousness. Parameters for infectivity, exposed time and recovery time can be adjusted based on the disease being modeled. Our society model is specific using the data described in the previous section to create a realistic synthetic population for a town in Ireland. The model includes transportation with agents moving between their current location and desired destination in a straight line in steps allowing them to interact with other agents along their route. The model environment includes maps of the towns being modelled. The small area boundary files from the CSO are use to determine the layout of the towns and zoning data is used to determine where residential, community and industrial areas should be placed.

The following sections provide a detailed description of the model. We use the ODD format for our model description (Grimm et al., 2010). It is a standard format used to describe agent-based models. The model description is for the model for an Irish town. Some of the following sections are include repetition from the main body of the thesis but are included here for completeness.

Purpose

The purpose of the model is to use openly available data to create an agent-based model of an infectious disease spread in an Irish town. The model is designed to be transferable to other towns. Specifically, we want to identify factors that might result in towns being more or less susceptible to an outbreak, in particular town density, town population and their interactions along with vaccination rates and socioeconomic makeup of the town.

Entities, state variables and scales

- **Agents/Individuals:** The model has one type of agent. The agents represent people in the town being modeled. The state variables for each agent including characteristics such as age, gender and economic status. Table A.1 shows the full list of variables for each agent.

- **Spatial Units:** Each grid cell or patch in Netlogo represents approximately 111 m^2 of land. The state variables for each grid cell within the model can be found in Table A.2.

- **Environment:** The model environment is made using data to determine the town boundaries, land use. For town boundaries we use data from the Irish Central Statistics Office (CSO). For accuracy the towns are broken up into small areas. A small area is a technical term used by the CSO to describe the smallest area over which census data is aggregated. A town with small area boundaries can be seen in Figure A.1. The white boundaries are the

State Variables	Description
Who	Unique Agent Id
Age	0 - 95
Familyid	Connects family members
Small_area	Region of town agent lives in
Adult?	Is the agent an adult. True/False
Home_patch	The coordinates where the agent lives
HH_type	Single, couple, couple with children, single father with children, single mother with children, couple plus others, couple with children plus others, single father with children plus others, single mother with children plus others, multi-family, other
HHSize	Number of agents in household. 1+
Couple	Is the agent part of a couple? True/False
Children?	Does the agent have children? True/False
Single_mom	Is the agent a single mother. True/False
Single _dad	Is the agent a single father. True/False
Child_type	U15(under 15), O15(over 15), UO15 (under and over 15
Child_size	Number of Children
Infant	Is the child an infant. True/False
Econ_stat	Economic Status of the agent. Work, student, Retired, Unemployed, Looking for First Job, Stay-at-home, Sick/Disabled
Job_type	Type of job the adult agent has. Professional Workers, Managerial and Technical, Non-Manual, Skilled Manual, Semi-Skilled, Unskilled
Work-patch	Coordinates of the agent's workplace
Sick?	Is the agent sick? True/False
Immune?	Is the agent immune? True/False
Exposed?	Is the agent exposed? True/False
Dayssick	Number of days the agent has been sick
Daysexposed	Number of days the agent has been exposed
Tickexposed	The time the agent was first exposed
Ticksexposed	The length of time the agent has been exposed
Ticksick	The time the agent first moved from exposed to sick
Ticksexposed	The length of time the agent has been sick
Immunity	The immunity the agent has to the disease. Values between 0 and 1
Dest-patch	The location the agent is moving towards
Num_contacts	List of the number of contacts an agent had had each time step
Avg_contacts	The average number of contacts the agent has had
Cont_table	Table with WHO numbers of other agents the agent has come into contact with during the time step
My_contacts_table	Table of number of contacts an agent has for each time step

Table A.1: State Variables for agents in the town model

State Variables	Description
Total	The total number of agents in the small area the patch is part of
Use	The use of the patch: home, work, school, open, town center, community, residential, commercial, mixed
Family	Does a family live here? True/False
House_type	What is the type of household on the patch? Single, couple, couple with children, single father with children, single mother with children, couple plus others, couple with children plus others, single father with children plus others, single mother with children plus others, multi-family, other
House-size	Size of the family on the patch
Work-size	Number of agents at the work place
Child_age	Age of children in family on the patch if the patch has a use of home. U15,O15, UO15

Table A.2: State Variables for each grid cell in the town model

boundaries of the small areas.

Environmental variables within the model include time, day of the week and

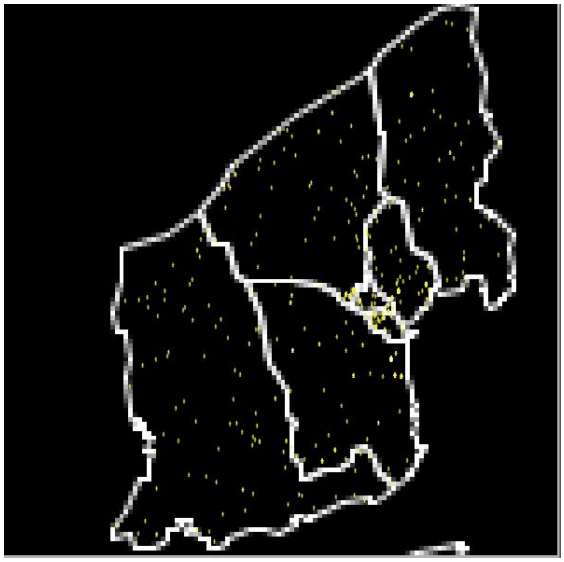


Figure A.1: Image of Schull, Ireland. The white boarders are the boarders of the small areas that make up the town.

week number. Each time step in the model represents two hours in a day. Seven days will make up one week. The first week in the year is considered week 1. Weeks are tracked to take into account summer vacation for students. Agents determine where they are moving based on the time. The model is affected by the day of the week, as agents will act differently on a weekend versus a weekday. The week number also affects the model; students will not attend school in the summer and will treat everyday as a weekend.

Process overview and scheduling

The model proceeds in discrete time steps that represent two hour. The model runs until there are no longer any agents who are exposed or immune. Each time step the following submodels are run: clock, move, infect, recover, update-global-variables, update-numcontacts, reset-contacts. Submodels are described in

Appendix A: submodels.

Design Concepts

- **Basic Principles:** We base the infection part of the model on an SEIR (susceptible, exposed, infected and recovered). It is a model that is widely used within infectious disease modelling. The idea is that when a susceptible agent come into contact with an infected agent there is a certain probability that the agent will become exposed to the disease. This probability is determined using values for R_0 for the disease, basic reproductive number and is defined as the expected number of individuals infected by one infectious individual in a completely susceptible population. It can be broken down into three components, number of contacts per unit time (c), the transmission probability per contact (p) and the duration of the infectiousness (d). The relationship can be seen in Equation A.1 (Thomas & Weber, 2001).

$$p = \frac{R_0}{cd} \tag{A.1}$$

Of the four variables in the equation, R_0, c, p, and d, three are known or can be estimated from our model, allowing us to determine the value for p. The agent will then progress from exposed to infected and finally to recovered. Agents will come into contact with other agents based on how they move through the environment. The model takes a simplistic approach to agent movement, with agents moving in a straight line between home

and workplaces or schools. Agents who do not have a workplace will move randomly in the town to locations in the town center, residential areas or recreational areas. During weekends and summer holidays for students all agents move randomly within the town.

- **Emergence:** The emerging result from the model is the course that the infection takes. Based on the type of agent that is initially infectious, the other agents that come into contact with the infectious agent and how long the contact lasts, patterns can emerge for how an outbreak will spread. For example, if a student is infected compared with an unemployed agent, the student will likely come into contact with more susceptible agents every day it attends school leading to a larger outbreak. Agents decisions to stay home when sick can also have a major effect on how the outbreak occurs.

- **Adaptation:** The current version of the model has little adaptation involved. Agents reproduce observed behaviours based on a set of rules given to them. For example, on weekends agents will move from one location to another a certain percentage of the time. If an agent becomes sick they will adapt their behaviour in that they can decide to go about their day as normal or to stay home.

- **Sensing:** As they move through the environment infected agents will sense if other agents that are close to them are susceptible.

- **Interaction:** The model assumes direct interaction between agents. If two

agents occupy the same space (both agents are on the same Netlogo patch) it is assumed that they have had some sort of interaction which may lead to the infection of an agent if one is susceptible and one is infected.

- **Stochasticity:** Agent movements in the model are partially random. Weekend movement for all agents, summer movement for students and everyday movements for unemployed, stay at home, retired, and sick and disabled agents are all determined stochastically. Agents will stay where they are x% of the time and (1- x)% of the time they will move to a randomly selected location with a given land use (town center, recreational or residential).

  Additionally stochasticity in the model is seen in the spread of the infectious disease through the population. When an infectious agent comes into contact with a susceptible agent there is a certain probability that determines if the susceptible agent will become exposed. Once exposed the length of time that agent will remain exposed before becoming infectious is determined by a probability distribution. Similarly the length of time an agent stays infectious is determined by a probability distribution.

- **Observation:** Every run of the model data is collected on the number of agents who are susceptible, exposed, infected and recovered at each time step. The output is collected at every time step in order to see how the infection changes over time. Data is also collected on the age of individuals infected along with their work status (student, work, unemployed etc.). The age and work status is only collected the time step that an agent becomes

infected. Finally the average number of contacts across all agents in the model is collected.

Initialization

The town was setup using the small area datasets from the Irish Central Statistics Office (CSO). Small areas are the smallest level of geography that the census data is counted for and are made up of between 50 and 200 homes. Small areas for each town are loaded into the model to determine the boundaries of the town. Zoning data is used to assign patches to a use of open space, town center, community, residential, commercial or mixed. The number of occupied households in each small area can be determined using household tables from the CSO. This number is then used to randomly generate the correct number of household within each small area and households are placed in residential patches. Patches with a zone of commercial, town center or mixed can be a workplace. School location datasets are used to find the locations of the schools in the town. Agents are added to the town based on the closest census data to the year being modelled. Census data is produced every five years with the two most recent datasets for the 2011 and 2016 census. If we were modelling an outbreak from 2012 the 2011 census data would be used. However, if the outbreak was in 2014, the 2016 census data would be used. The number of agents in the simulation will be the number of people living in the town. The following steps are used in populating the town with agents and are performed for each individual small area:

- Each household is assigned a type (single, couple, couple plus others, couple with children, couple with children plus others, single parent, single parent plus others or other).
- Adults are added into each household. One agent is added to households with types single, single parent and other. Two agents are added to the households with type couple.
- Adults in each household are assigned a sex and age based on a probability distribution determined from the CSO census age, sex tables for the relevant small area.
  - The age categories provided by the CSO are by year until 19 after which ages are reported in ranges of five years, for example ages 20-24 or ages 60-64, and then anyone over 85 is combined into one age bracket. To have all ages represented in our model in each age bracket we randomly assign individuals one of the five ages represented in that bracket.
  - Couples are assigned opposite genders and ages within 10 years of each other.
- If a household type includes children a probability distribution determined from the relevant census data is used to determine if all children in the house are under 15, over 15, or both over and under 15.
- Data from the family units with children by size and age of children table is used to determine the probability that each household with type child has

1, 2, 3, 4 or 5 children in the household.

- Children are added into each household.
- Children are assigned a sex and age based on a probability distribution extracted from the relevant census data and the type of children the household is assigned to (under 15, over 15 or both under and over 15).
- If the total number of agents populating a small area is not equal to the total number of agents who should be in the area based on the CSO data, additional agents are added and randomly assigned to households of types couple plus others, couple with children plus others, single parent plus others or other.
- All agents are assigned an economic status based on CSO data.
  - Agents over 65 are assigned to retired.
  - Agents between the ages of 5 and 14 are assigned to student.
  - Agents between the ages of 15 and 18 are first assigned to student. If there are more agents aged 15 to 18 in the small area than the number of students in the same age categories then the agents are assigned to looking for first job. If there are still more agents aged 15 to 18 they are then assigned an economic status of work, unemployed or sick/ disabled following the distribution for these categories for the relevant small area.
- Adult agents under 65 are assigned to work, looking for first job, unemployed,

sick/ disabled or stay at home following the distribution for these categories for the relevant small area. Agents are only assigned to stay at home if they are part of a couple.

- Agents under the age of 5 are assigned to student if they have no stay at home parent. If they have a stay at home parent then a probability determines if the agent will be assigned to student.
- Agents with an economic status of work are randomly assigned to one of the work-places in town.
- Agents can be assigned a social class.
  - Agents with an economic status of work are given one of the following social classes: Professional Workers, Managerial and Technical, Non-Manual,Skilled Manual, Semi-Skilled, or Unskilled.
  - Agents who are retired are given a social class of retired and agents who are unemployed, sick/ disabled, looking for first job, or stay-at-home are given the social class other.
  - Agents who are younger than 18 are not given a social class.
  - Households are assigned a household social class. This is randomly selected from one of the adults in the household. All agents in a household, including children, will have the same household social class.
- If vaccinations are included in the model agents are given an immunity level based on the disease being modeled and vaccination data for that disease.

Irish vaccination data is used to determine the percentage of each age group that have received vaccinations for the infectious disease being modelled. For example, if 90% of 1 year olds in Ireland had been given the MMR vaccination in 2011 and we are running a model for 2012, we give each agent in the model with an age of 2 a 90% chance of having been vaccinated. If an agent is vaccinated they are given a 97% chance of being immune to the disease. This takes into account vaccination failure and is based on the vaccine effectiveness rate for measles (Nelson & Williams, 2007). Half of the agents with age less than 1 are given immunity to a disease to mimic passive immunity infants receive from their mothers (Nicoara et al., 1999). For any agents that have an age corresponding to a vaccination year not in our data we give a 99% chance of being immune. Prior to vaccination campaigns the majority of the population would have either had or been exposed to childhood diseases, such as measles, leaving them immune in later life. If socioeconomic status is used in the model, the percent of individuals receiving vaccinations is adjusted based on the household socioeconomic status of the individual.

- Finally, a given number of agents are given the status of infected.

Segregation Burn-in Model

In order to account for clustering in neighborhoods by socioeconomic status we can add an additional step into our model setup. Once the model is populated

with agents and all of the agents are assigned the appropriate characteristics we give households the opportunity to move using a Schelling type segregation model (Schelling, 1969). Segregation models, specifically Schelling models show how a small individual preference to not be a racial minority in a neighbourhood leads to neighbourhood segregation. However, unlike in Schelling's models where race is used as the segregating factor we use social class.

The segregation model is run on time steps and a tolerance and a radius are set for the model. The radius is how far away the agents will look for similar households. The tolerance determines what percent of similar households within the radius the agents desire. During the segregation process at each time step each household does the following

- One of the adult agents calculate the household utility by finding the proportion of households within the radius that have the same family social class.
- If the household utility is greater then or equal to the tolerance the household will not move.
- If the utility is less than the tolerance the agent will search for empty households in the small area.
- For each empty household the agent will determine what their utility would be at that household.
- The household will then move to the empty household that will give them

them the greatest utility.

  - If no households will give the agents a greater utility then their current household they do not move.
  - If there are no other households within the radius of the agents they will set their utility to the tolerance and will not move.
  - Each household gets the opportunity to move once each time step.

After a full time step of no households moving the segregation model stops. Once the segregation model has stopped running, agents who are students are assigned to a school

- Agents aged 13 and up with an economic status of student are assigned to the secondary school closest to their home patch.
- Agents aged 4 to 12 with an economic status of student are assigned to a primary school closest to their home patch.
- Agents aged 3 and below with an economic status of student are assigned to a pre-school closest to their home patch.

Input Data

The model does not use input data to represent time-varying processes.

Submodels

- **Clock:** The clock submodel keeps track of the time, day and week of the model. The time is the determined as the modulus of ticks and 12. When the time goes back to 0 the day is increased by one. If the day is 8, the week is increased by one and the day goes back to 1.

- **Move:** In the model agents use straight-line transportation. Agents will move between their home and destination in a straight line following the most direct route. Although this is a naive model, for small towns where distances travelled are short, such as those discussed in this paper, it is effective.

  - Agents who are working leave their home on the fourth time step of the day, which would be equivalent to between 8am and 9am, arrive at work over one time step, spend 4 time steps (8 hours) at work and then return home.

  - Students also leave on the fourth time step but only spend 3 time steps (6 hours) at school.

  - Stay at home agents who have children in primary school travel with their children to school on the 4th time step and then return home during the same time step. Between the fourth and seventh time step (when students return home from school) stay at home agents move randomly throughout the town: at each step if an agent is at home

they have a 50% chance of staying at home. If not at home, an agent has a 50% chance of picking a new destination in town and moving there. At the 7th time step of the day the stay at home agents will go to their child's school and then travel home.

- Stay at home agents who do not have a child in primary school move randomly throughout the town between the fourth time step and the seventh time step the same way stay at home parents move when they are not travelling with their children to school or home.
- Agents younger than 4 who are not assigned to a preschool move with their stay at home parent throughout the day.
- Agents who are unemployed, looking for their first job, retired or sick/ disabled move randomly throughout the town between the 4th and 10th time steps of the day.
- If an agent is infected with the disease simulated within a model then their behaviour is affected. Infected agents have a certain probability of staying home. If they are working, the agents will stay home 30% of the time. Students will stay home 70% of the time. Unemployed agents will stay home 75% of the time, and stay at home agents with primary school children will stay at home 10% of the time when accompanying children to school and 50% of the time when moving around town. Stay at home agents with non-primary schoolchildren will stay at home 50% of the time.

  - All agents will move randomly through the town on the weekends.

- **Infect:** When an infected agent comes into contact with a susceptible agent, the infected agent will determine if they will infect the susceptible agent based on the infection rate, a variable chosen at the start of the simulation. If the infected agent determines it will infect the susceptible agent, the susceptible agent will change their health status from susceptible to exposed.

- **Recover:** An exposed agent will use a probability distribution to determine the number of time steps it will stay exposed before it becomes infected. Similarly, when the agent switches from exposed to infected the agent will use a probability distribution to determine the number of time steps before they are recovered/immune. Once an agent has recovered they cannot become infected again.

- **Update Global Variables:** At the end of each time step, all global variables are updated. The counts and percent of susceptible, exposed, infected and recovered agents are all calculated. The average number of contacts across all agents in the model is calculated by taking the average of each agent's contacts.

- **Update Numcontacts:** As agents move through the environment they keep track of every agent, they come into contact with (agents on the same patch of the environment). At the end of the time step, the agents will calculate the number of unique contacts they have had and calculate an

average number of contacts they have had during the simulation. An overall average for number of contacts across all agents is then taken.

- **Reset Contacts:** After the average number of contacts has been calculated, the contacts each agent has had is reset to 0 so that they can calculate a new average for the next tick.

Appendix B

County Model Description

Model Description

The model we describe in this section is a model for measles outbreaks within an Irish county. It is presented in Chapter 9 of this thesis. The same as the model described in Appendix A, we use the computer software Netlogo (Wilensky, 1999a) to implement our model. This model aims to scale up the model by making some assumptions about the level of detail necessary to produce accurate results.

The basis of the model is the same as that for the model in Appendix A: our model is a data driven agent-based model for the human airborne infectious diseases measles. It follows an SEIR (susceptible, exposed, infected, and recovered) type compartmental model with the agents moving between the four state relating to infectiousness. Our society model is specific using the data to create a realistic synthetic population for a county in Ireland. The model includes transportation with agents moving between their current location and desired destination using

predetermined destinations or destinations selected using a gravity model. The model environment includes the small areas from the CSO which agents can move between.

The following sections provide a detailed description of the model. We use the ODD format for our model description (Grimm et al., 2010). Similar to Appendix A this appendix includes repetition from the main body of the thesis and the previous appendix but we include this for completeness of the model description.

Purpose

The purpose of the model is to use openly available data to create an agent-based model of an infectious disease spread in a region specifically an Irish county. The model is designed to be transferable to other regions and counties. We aim to identify factors that might result in the region or towns within the region being more or less susceptible to an outbreak.

Entities, state variables and scales

- **Agents/Individuals:** The model has one type of agent. The agents represent people in the town being modeled. The state variables for each agent including characteristics such as age, gender and economic status. Table B.1 shows the full list of variables for each agent.

- **Spatial Units:** Each grid cell or patch in Netlogo represents a single small area. A small area is a geographic census area defined by the CSO as having

State Variables	Description
Who	Unique Agent Id
Age	0 - 95
Familyid	Connects family members
Small_area	Region of town agent lives in
Adult?	Is the agent an adult. True/False
Home_patch	The coordinates where the agent lives
Children?	Does the agent have children? True/False
Infant	Is the child an infant. True/False
Econ_stat	Economic Status of the agent. Work, student, Retired, Unemployed, Looking for First Job, Stay-at-home, Sick/Disabled
Work-patch	Coordinates of the agent's workplace
Workplace	ID of workplace
School	ID of school agent attends
family_network	Other agents in the agent's family
work_network	Other agents in the agent's workplace or school
class_network	Other agents in the agent's school who are the same age
location	Where is the agent in the small area Community, Work, School, Home
Sick?	Is the agent sick? True/False
Immune?	Is the agent immune? True/False
Exposed?	Is the agent exposed? True/False
Dayssick	Number of days the agent has been sick
Daysexposed	Number of days the agent has been exposed
Tickexposed	The time the agent was first exposed
Ticksexposed	The length of time the agent has been exposed
Ticksick	The time the agent first moved from exposed to sick
Ticksexposed	The length of time the agent has been sick
Where_sick	Where was the agent infected? Community, Work, School, Home
Immunity	The immunity the agent has to the disease. Values between 0 and 1
Dest-patch	The location the agent is moving towards
Num_contacts	List of the number of contacts an agent had had each time step
Avg_contacts	The average number of contacts the agent has had
Cont_table	Table with WHO numbers of other agents the agent has come into contact with during the time step
My_contacts_table	Table of number of contacts an agent has for each time step

Table B.1: State Variables for agents in the county model

50 to 200 dwellings. It is the smallest geographic area over which the census is aggregated. The state variables for each grid cell within the model can be found in Table B.2.

State Variables	Description
Smarea	The small area ID for the grid cell from the CSO data
townname	Name of the town that the small area is part of
County	Name of the county
primarycount	Count of primary schools in the small area
secondarycount	Count of secondary schools in the small area
distances	List of real world distances to other small areas
otherpatches	List of the small area IDs of the other grid cells
moveprob	The probability that the agents on the cell will move to another cell based on the distance
primarydist	List of distances to all primary schools in the model
secondarydist	List of distance to all secondary schools in the model

Table B.2: State Variables for each grid cell in the county model

- **Environment:** The Netlogo world is a two dimensional grid where the squares that make up the grid are referred to as patches. Patches in our model represent the small areas within a county. All agents that are in a small area at a given time are physically in the same location. However, agents will keep track of their location within that small area. There are four possibilities for agents locations within a small area: home, work, school, or the community. Who an agent comes into contact with depends on that location. For example, if an agent is at home, they know they are at home and will only come into contact with other members of their household who are also at home. Agents in the community within a small area will come into contact with other agents in that small area in the community but will not come into contact with all agents in the community. If two agents from the same household are in the community there is a larger probability of them coming into contact then two agents who are in the same workplace or school who in turn have a larger probability of coming into contact than two agents who have no other connection. All agents within a small area

patch have access to information about the patch they are in, including the number of primary and secondary schools in the small area. As well as the real world distances between the center of that small area and all other small areas in the model.

Environmental variables within the model include time, day of the week and week number. Each time step in the model represents two hours in a day. Seven days will make up one week. The first week in the year is considered week 1. Weeks are tracked to take into account summer vacation for students. Agents determine where they are moving based on the time. The model is affected by the day of the week, as agents will act differently on a weekend versus a weekday. The week number also affects the model; students will not attend school in the summer and will treat everyday as a weekend.

Process overview and scheduling

The model proceeds in discrete time steps that represent two hour. The model runs until there are no longer any agents who are exposed or infected. Each time step the following submodels are run: clock, move, infect, recover, update-global-variables, find-contacts, and who-sick. Submodels are described in Appendix B: submodels.

Design Concepts

- **Basic Principles:** We base the infection part of the model on an SEIR (susceptible, exposed, infected and recovered). It is a model that is widely

used within infectious disease modelling. The idea is that when a susceptible agent come into contact with an infected agent there is a certain probability that the agent will become exposed to the disease. This probability is determined using values for R_0 for the disease, basic reproductive number and is defined as the expected number of individuals infected by one infectious individual in a completely susceptible population. It can be broken down into three components, number of contacts per unit time (c), the transmission probability per contact (p) and the duration of the infectiousness (d). The relationship can be seen in Equation B.1 (Thomas & Weber, 2001).

$$p = \frac{R_0}{cd} \tag{B.1}$$

Of the four variables in the equation, R_0, c, p, and d, three are known or can be estimated from our model, allowing us to determine the value for p. The agent will then progress from exposed to infected and finally to recovered. Agents will come into contact with other agents based on how they move through the environment.

The model uses a gravity model to agent movement. Agents movements that are not predetermined (moving to home, school or work at given times) will be determined based on an inverse relationship with the distance to the location and proportionally to the population density of the location. Agents are pulled to locations where there are already a lot of other agents

and pushed away from locations that are farther away. During weekends and summer holidays all agents who are students move using the gravity model.

- **Emergence:** The emerging result from the model is the course that the infection takes. Based on the type of agent that is initially infectious, the other agents that come into contact with the infectious agent and how long the contact lasts, patterns can emerge for how an outbreak will spread. For example, if a student is infected compared with an unemployed agent, the student will likely come into contact with more susceptible agents every day it attends school leading to a larger outbreak. Agents decisions to stay home when sick can also have a major effect on how the outbreak occurs.

- **Adaptation:** The current version of the model has little adaptation involved. Agents reproduce observed behaviours based on a set of rules given to them. For example, on weekends agents will move from one location to another a certain percentage of the time. If an agent becomes sick they will adapt their behaviour in that they can decide to go about their day as normal or to stay home.

- **Sensing:** As they move through the environment infected agents will sense if other agents that are close to them are susceptible.

- **Interaction:** The model assumes direct interaction between agents. If two agents occupy the same space (both agents are on the same Netlogo patch) they may be in contact with each other. If they are in contact is determined

by their location in the small area (home, work, school or community) and if the other agent is in their network. An interaction may lead to the infection of an agent if one is susceptible and one is infected.

- **Stochasticity:** Agent movements in the model are partially random. Weekend movement for all agents, summer movement for students and everyday movements for unemployed, stay at home, retired, and sick and disabled agents are all determined stochastically. Agents will stay where they are x% of the time and (1- x)% of the time they will move to a selected location selected using a gravity model.

  Additionally stochasticity in the model is seen in the spread of the infectious disease through the population. When an infectious agent comes into contact with a susceptible agent there is a certain probability that determines if the susceptible agent will become exposed. Once exposed the length of time that agent will remain exposed before becoming infectious is determined by a probability distribution. Similarly the length of time an agent stays infectious is determined by a probability distribution.

- **Observation:** Every run of the model data is collected on the number of agents who are susceptible, exposed, infected and recovered at each time step. The output is collected at every time step in order to see how the infection changes over time. Data is also collected on the age of individuals infected along with their home location (what small area they live in). The age and home location is only collected the time step that an agent becomes

infected. Finally the average number of contacts across all agents in the model is collected.

Initialization

The initiallization process is only done once for each county. After it is run once the *world* in Netlogo is saved and is reloaded in for each run. The county was setup using the small area datasets from the Irish Central Statistics Office (CSO). Small areas are the smallest level of geography that the census data is counted for and are made up of between 50 and 200 homes. Small areas for each town are loaded into the model with a patch representing a small area. School location datasets are used to find the small areas that schools are located in within the county. Each small area patch records the real world distance between the center of the small area and the center of every other small area in the model. Agents are added to the town based on the closest census data to the year being modelled. Census data is produced every five years with the two most recent datasets for the 2011 and 2016 census. If we were modelling an outbreak from 2012 the 2011 census data would be used. However, if the outbreak was in 2014, the 2016 census data would be used. The number of agents in the simulation will be the number of people living in the town. The following steps are used in populating the town with agents and are performed for each individual small area:

- Each household is assigned a type (single, couple, couple plus others, couple with children, couple with children plus others, single parent, single par-

ent plus others or other) and a family id that links all agents in the same household.

- Adults are added into each household. One agent is added to households with types single, single parent and other. Two agents are added to the households with type couple.

- Adults in each household are assigned a sex and age based on a probability distribution determined from the CSO census age, sex tables for the relevant small area.

  - The age categories provided by the CSO are by year until 19 after which ages are reported in ranges of five years, for example ages 20-24 or ages 60-64, and then anyone over 85 is combined into one age bracket. To have all ages represented in our model in each age bracket we randomly assign individuals one of the five ages represented in that bracket.

  - Couples are assigned opposite genders and ages within 10 years of each other.

- If a household type includes children a probability distribution determined from the relevant census data is used to determine if all children in the house are under 15, over 15, or both over and under 15.

- Data from the family units with children by size and age of children table is used to determine the probability that each household with type child has 1, 2, 3, 4 or 5 children in the household.

- Children are added into each household.
- Children are assigned a sex and age based on a probability distribution extracted from the relevant census data and the type of children the household is assigned to (under 15, over 15 or both under and over 15).
- If the total number of agents populating a small area is not equal to the total number of agents who should be in the area based on the CSO data, additional agents are added and randomly assigned to households of types couple plus others, couple with children plus others, single parent plus others or other.
- All agents are assigned an economic status based on CSO data.
  - Agents over 65 are assigned to retired.
  - Agents between the ages of 5 and 14 are assigned to student.
  - Agents between the ages of 15 and 18 are first assigned to student. If there are more agents aged 15 to 18 in the small area than the number of students in the same age categories then the agents are assigned to looking for first job. If there are still more agents aged 15 to 18 they are then assigned an economic status of work, unemployed or sick/ disabled following the distribution for these categories for the relevant small area.
  - Adult agents under 65 are assigned to work, looking for first job, unemployed, sick/ disabled or stay at home following the distribution for

these categories for the relevant small area. Agents are only assigned to stay at home if they are part of a couple.

  - Agents under the age of 5 are assigned to student if they have no stay at home parent. If they have a stay at home parent then a probability determines if the agent will be assigned to student.

- Agents with an economic status of work are assigned to a workplace based on commuting data. The data is used to determine the probability that an agent in a given electoral division (CSO area one step above small areas) will commute to any other electoral division. The commuting data also includes agents who commute out of the region, if an agent ends up commuting out of the region they will select a patch in the model that is not designated as a small area and commute there during work hours.

- Agents aged 13 and up with an economic status of student are assigned to a secondary school based on a distance function. The closer the secondary school to the student the greater the probability the student will attend the school.

- Agents aged 4 to 12 with an economic status of student are assigned to a primary school based on a distance function. The closer the primary school to the student the greater the probability they student will attend the school.

- Agents aged 3 and below with an economic status of student are assigned to a pre-school closest to their home patch.

- When all agents in the region are assigned a workplace or school, social networks are created.
  - Agents create a family network with all other agents in their household.
  - Agents create a work network with all other agents that work in the same workplace.
  - Students create a school network with all other agents in their school and a class network with other agents in the school who are the same age as them.
- If vaccinations are included in the model agents are given an immunity level based on the disease being modeled and vaccination data for that disease. Irish vaccination data is used to determine the percentage of each age group that have received vaccinations for the infectious disease being modelled. For example, if 90% of 1 year olds in Ireland had been given the MMR vaccination in 2011 and we are running a model for 2012, we give each agent in the model with an age of 2 a 90% chance of having been vaccinated. If an agent is vaccinated they are given a 97% chance of being immune to the disease. This takes into account vaccination failure and is based on the vaccine effectiveness rate for measles (Nelson & Williams, 2007). Half of the agents with age less than 1 are given immunity to a disease to mimic passive immunity infants receive from their mothers (Nicoara et al., 1999). For any agents that have an age corresponding to a vaccination year not in our data we give a 99% chance of being immune. Prior to vaccination campaigns

the majority of the population would have either had or been exposed to childhood diseases, such as measles, leaving them immune in later life. If socioeconomic status is used in the model, the percent of individuals receiving vaccinations is adjusted based on the household socioeconomic status of the individual.

- Finally, a given number of agents are given the status of infected.

Input Data

The model does not use input data to represent time-varying processes.

Submodels

- **Clock:** The clock submodel keeps track of the time, day and week of the model. The time is the determined as the modulus of ticks and 12. When the time goes back to 0 the day is increased by one. If the day is 8, the week is increased by one and the day goes back to 1.

- **Move:** In the model agents move in one step between their current location and their desired destination. Some movements are predetermined by the model rules others are determined using a gravity model. A gravity model is used to determine agents movements throughout the region. This is an alternative to the simple random movements used in earlier models. Random selection of the agents next destination is a naive model that was acceptable under the assumptions of a small town or city but when the area over which

agents can travel increases to say a county this assumption of random movement does not hold as well. For example, in a small town moving to the other side of the town is much more likely than moving from one side of the county to the other. To account for this a gravity model is used to determine movements. Gravity models are a type of transportation model that is similar in formula to Newton's gravitation model. A traditional gravity model gives the interactions between two location pairs and determines those interactions based on the characteristics of a location and the distance between locations (Rodrigue et al., 2006). In the model, agents move between home and school or work at certain predetermined times and will return home at predetermined times. On weekends, summers for students and after school or work hours agents will move through the community, these movements are determined by the gravity model. The probability of an agent moving to another small area is proportional to the population density of the small area, an area with more agents is more attractive, and inversely proportional to the distance to the small area from the agents current location, areas that are farther away are less attractive. We feel that this transportation model provides a more accurate model of movement within a larger area than that in the original town model.

The set predetermined movements are as follows:

- Agents who are working leave their home on the fourth time step of the day, which would be equivalent to between 8am and 9am, arrive at

work over one time step, spend 4 time steps (8 hours) at work and then return home.

- Students also leave on the fourth time step but only spend 3 time steps (6 hours) at school.
- Stay at home agents who have children in primary school travel with their children to school on the 4th time step and then return home during the same time step. Between the fourth and seventh time step (when students return home from school) stay at home agents move throughout the town: at each step if an agent is at home they have a 50% chance of staying at home. If not at home, an agent has a 50% chance of picking a new destination in town with the gravity model and moving there. At the 7th time step of the day the stay at home agents will go to their child's school and then travel home.
- Stay at home agents who do not have a child in primary school move throughout the town using the gravity model between the fourth time step and the seventh time step the same way stay at home parents move when they are not travelling with their children to school or home.
- Agents younger than 4 who are not assigned to a preschool move with their stay at home parent throughout the day.
- Agents who are unemployed, looking for their first job, retired or sick/ disabled move throughout the town with the gravity model between the 4th and 10th time steps of the day.

- If an agent is infected with the disease simulated within a model then their behaviour is affected. Infected agents have a certain probability of staying home. If they are working, the agents will stay home 30% of the time. Students will stay home 70% of the time. Unemployed agents will stay home 75% of the time, and stay at home agents with primary school children will stay at home 10% of the time when accompanying children to school and 50% of the time when moving around town. Stay at home agents with non-primary schoolchildren will stay at home 50% of the time.
- All agents will move throughout the town on the weekends using the gravity model.
- After each move agents will make note of what other agents they have come into contact with using their social networks.
    * If an agent is at home they will come into contact with all other agents in their family network who are also at home.
    * If an agent is at school they will come into contact with other agents also in school who are in their class network a given percentage of the time and with other agents also in their school who are in their school network a given percentage of the time. (The user of the model sets the percentage chance of contact but the chance that an agent comes into contact with another agent in their class network should be higher than the chance they come into contact with a

member of their school network.)

    - If an agent is at work they will come into contact with other agents in their work network a given percent of the time.
    - If an agent is in the community they will come into contact with other agents in the same small area also in the community a given percent of the time. This percent is determined by the social networks the agents are in. There can be a different chance for agents in the same family network, class network, school network, work network and agents who are not in each others social networks this is set by the user.

- **Infect:** When an infected agent comes into contact with a susceptible agent, the infected agent will determine if they will infect the susceptible agent based on the infection rate, a variable chosen at the start of the simulation. If the infected agent determines it will infect the susceptible agent, the susceptible agent will change their health status from susceptible to exposed.

- **Recover:** An exposed agent will use a probability distribution to determine the number of time steps it will stay exposed before it becomes infected. Similarly, when the agent switches from exposed to infected the agent will use a probability distribution to determine the number of time steps before they are recovered/immune. Once an agent has recovered they cannot become infected again.

- **Update Global Variables:** At the end of each time step, all global variables are updated. The counts and percent of susceptible, exposed, infected and recovered agents are all calculated. The average number of contacts across all agents in the model is calculated by taking the average of each agent's contacts.

- **Find-Contacts:** If the user has selected to keep track of contacts, find contacts will run each each step. Each agent has a vector that has recorded the number of other agents they have come into contact with each time step. Find-contacts has each agent find the average total contacts and add this to a list of average contacts for all agents. The average across all agents is then taken to find the average number of contacts by an agent each time step.

- **Who-sick:** A sub model to find and report the ID, age and home location of the agents who became infected in a given time step.

Appendix C

Hybrid Model Description

Model Description

The model described in this Appendix is the hybrid agent-based and equation based model presented in Chapter 10. We use the computer software Netlogo (Wilensky, 1999a) to implement our model. Netlogo is an easy to use and popular environment for creating ABMs (Gilbert, 2008). We chose to use Netlogo due to the increasing popularity of the platform with agent-based modellers and its ability as a medium to high/large scale modelling platform. It does, however, have disadvantages, one of the biggest being the speed of the program. When modelling simulations with a small number of agents Netlogo works well, however, once the number of agents gets large enough the simulation slows down.

Our model is hybrid agent-based and equation based model for human airborne infectious diseases measles. It follows an SEIR (susceptible, exposed, infected, and recovered) type compartmental model with the agents moving between the

four state relating to infectiousness. However, the disease model can switch back and forth between agent-based and equation based depending on the number of infected agents. Our society model is specific using the data to create a realistic synthetic population for a county in Ireland. The model includes transportation with agents moving between their current location and desired destination using predetermined destinations or destinations selected using a gravity model. The model environment includes the small areas from the CSO which agents can move between.

The following sections provide a detailed description of the model. We use the ODD format for our model description (Grimm et al., 2010). There is repetition in this appendix especially with Appendix B and Chapter 10. We choose to include the repetition for completeness of the model description.

Purpose

The purpose of the model is to use openly available data to create a hybrid agent-based and equation based model of an infectious disease spread in a region specifically an Irish county. The model is designed to be transferable to other regions and counties. We want to identify factors that might result in the region or towns within the region being more or less susceptible to an outbreak and to test various intervention strategies.

Entities, state variables and scales

- **Agents/Individuals:** The model has one type of agent. The agents represent people in the town being modeled. The state variables for each agent including characteristics such as age, gender and economic status. Table C.1 in Appendix C shows the full list of variables for each agent.

- **Spatial Units:** Each grid cell or patch in Netlogo represents a single small area. A small area is a geographic census area defined by the CSO as having 50 to 200 dwellings. It is the smallest geographic area over which the census is aggregated. The state variables for each grid cell within the model can be found in Table C.2 in Appendix C.

- **Environment:** The Netlogo world is a two dimensional grid where the squares that make up the grid are referred to as patches. Patches in our model represent the small areas within a county. All agents that are in a small area at a given time are physically in the same location. However, agents will keep track of their location within that small area. There are four possibilities for agents locations within a small area: home, work, school, or the community. Who an agent comes into contact with depends on that location. For example, if an agent is at home, they know they are at home and will only come into contact with other members of their household who are also at home. Agents in the community within a small area will come into contact with other agents in that small area in the community but will

State Variables	Description
Who	Unique Agent Id
Age	0 - 95
Familyid	Connects family members
Small_area	Region of town agent lives in
Adult?	Is the agent an adult. True/False
Home_patch	The coordinates where the agent lives
Children?	Does the agent have children? True/False
Infant	Is the child an infant. True/False
Econ_stat	Economic Status of the agent. Work, student, Retired, Unemployed, Looking for First Job, Stay-at-home, Sick/Disabled
Work-patch	Coordinates of the agent's workplace
Workplace	ID of workplace
School	ID of school agent attends
family_network	Other agents in the agent's family
work_network	Other agents in the agent's workplace or school
class_network	Other agents in the agent's school who are the same age
location	Where is the agent in the small area Community, Work, School, Home
Sick?	Is the agent sick? True/False
Immune?	Is the agent immune? True/False
Exposed?	Is the agent exposed? True/False
Dayssick	Number of days the agent has been sick
Daysexposed	Number of days the agent has been exposed
Tickexposed	The time the agent was first exposed
Ticksexposed	The length of time the agent has been exposed
Ticksick	The time the agent first moved from exposed to sick
Ticksexposed	The length of time the agent has been sick
Where_sick	Where was the agent infected? Community, Work, School, Home
Immunity	The immunity the agent has to the disease. Values between 0 and 1
Dest-patch	The location the agent is moving towards
Num_contacts	List of the number of contacts an agent had had each time step
Avg_contacts	The average number of contacts the agent has had
Cont_table	Table with WHO numbers of other agents the agent has come into contact with during the time step
My_contacts_table	Table of number of contacts an agent has for each time step

Table C.1: State Variables for agents in the hybrid county model

not come into contact with all agents in the community. If two agents from the same household are in the community there is a larger probability of them coming into contact then two agents who are in the same workplace or school who in turn have a larger probability of coming into contact than

State Variables	Description
Smarea	The small area ID for the grid cell from the CSO data
townname	Name of the town that the small area is part of
County	Name of the county
primarycount	Count of primary schools in the small area
secondarycount	Count of secondary schools in the small area
distances	List of real world distances to other small areas
otherpatches	List of the small area IDs of the other grid cells
moveprob	The probability that the agents on the cell will move to another cell based on the distance
primarydist	List of distances to all primary schools in the model
secondarydist	List of distance to all secondary schools in the model
switched	Has the town the small area is in switched to equation based?
other_ed	List of other small areas in the town
turtsed	List of agents in the town
Si	Number of susceptible agents in the current time step
Ei	Number of exposed agents in the current time step
Ii	Number of infected agents in the current time step
Ri	Number of recovered agents in the current time step
Si1	Number of susceptible agents in the next time step
Ei1	Number of exposed agents in the next time step
Ii1	Number of infected agents in the next time step
Ri1	Number of recovered agents in the next time step

Table C.2: State Variables for each grid cell in the hybrid county model

two agents who have no other connection. All agents within a small area patch have access to information about the patch they are in, including the number of primary and secondary schools in the small area. As well as the real world distances between the center of that small area and all other small areas in the model.

Environmental variables within the model include time, day of the week and week number. Each time step in the model represents two hours in a day. Seven days will make up one week. The first week in the year is considered week 1. Weeks are tracked to take into account summer vacation for students. Agents determine where they are moving based on the time. The model is affected by the day of the week, as agents will act differently on a weekend

versus a weekday. The week number also affects the model; students will not attend school in the summer and will treat everyday as a weekend.

Process overview and scheduling

The model proceeds in discrete time steps that represent two hour. The model runs until there are no longer any agents who are exposed or infected. Each time step the following submodels are run: clock, move, infect, recover, update-global-variables, find-contacts, and who-sick. Submodels are described in Appendix C: Submodels.

Design Concepts

- **Basic Principles:** We base the infection part of the model on an SEIR (susceptible, exposed, infected and recovered). It is a model that is widely used within infectious disease modelling. The idea is that when a susceptible agent come into contact with an infected agent there is a certain probability that the agent will become exposed to the disease. This probability is determined using values for R_0 for the disease, basic reproductive number and is defined as the expected number of individuals infected by one infectious individual in a completely susceptible population. It can be broken down into three components, number of contacts per unit time (c), the transmission probability per contact (p) and the duration of the infectiousness (d). The relationship can be seen in Equation C.1 (Thomas & Weber, 2001).

$$p = \frac{R_0}{cd} \tag{C.1}$$

Of the four variables in the equation, R_0, c, p, and d, three are known or can be estimated from our model, allowing us to determine the value for p. The agent will then progress from exposed to infected and finally to recovered. Agents will come into contact with other agents based on how they move through the environment.

The model uses a gravity model to agent movement. Agents movements that are not predetermined (moving to home, school or work at given times) will be determined based on an inverse relationship with the distance to the location and proportionally to the population density of the location. Agents are pulled to locations where there are already a lot of other agents and pushed away from locations that are farther away During weekends and summer holidays agents who are students move using the gravity model.

We allow for the model to switch between an agent-based and equation based disease component and consider that the heterogeneous mixing and actions of the agents is most important when there are a small number of infected agents. At this point the actions of one agent can play a large part in how the outbreak unfolds or if an outbreak occurs at all. For example, if the initially infected agent stays home and does not come into contact with anyone else an outbreak will not occur. Thus at this stage it is important to have an

agent-based disease component. However, we propose that once there are enough infected agents the individual actions do not have as large of an impact on the outbreak and thus using an equation based model will not result in much loss of fidelity of the results.

- **Emergence:** The emerging result from the model is the course that the infection takes. Based on the type of agent that is initially infectious, the other agents that come into contact with the infectious agent and how long the contact lasts, patterns can emerge for how an outbreak will spread. For example, if a student is infected compared with an unemployed agent, the student will likely come into contact with more susceptible agents every day it attends school leading to a larger outbreak. Agents decisions to stay home when sick can also have a major effect on how the outbreak occurs.

- **Adaptation:** The current version of the model has little adaptation involved. Agents reproduce observed behaviours based on a set of rules given to them. For example, on weekends agents will move from one location to another a certain percentage of the time. If an agent becomes sick they will adapt their behaviour in that they can decide to go about their day as normal or to stay home.

- **Sensing:** As they move through the environment infected agents will sense if other agents that are close to them are susceptible.

- **Interaction:** The model assumes direct interaction between agents. If two

agents occupy the same space (both agents are on the same Netlogo patch) they may be in contact with each other. If are in contact is determined by their location in the small area (home, work, school or community) and if the other agent is in their network. An interaction may lead to the infection of an agent if one is susceptible and one is infected.

- **Stochasticity:** Agent movements in the model are partially random. Weekend movement for all agents, summer movement for students and everyday movements for unemployed, stay at home, retired, and sick and disabled agents are all determined stochastically. Agents will stay where they are x% of the time and (1- x)% of the time they will move to a selected location selected using a gravity model.

  Additionally stochasticity in the model is seen in the spread of the infectious disease through the population. When an infectious agent comes into contact with a susceptible agent there is a certain probability that determines if the susceptible agent will become exposed. Once exposed the length of time that an agent will remain exposed before becoming infectious is determined by a probability distribution. Similarly the length of time an agent stays infectious is determined by a probability distribution.

- **Observation:** Every run of the model data is collected on the number of agents who are susceptible, exposed, infected and recovered at each time step. The output is collected at every time step in order to see how the infection changes over time. Data is also collected on the age of individuals

infected along with their home location (what small area they live in). The age and home location is only collected the time step that an agent becomes infected. Finally the average number of contacts across all agents in the model is collected.

Initialization

The initiallization process is only done once for each county. After it is run once the *world* in Netlogo is saved and is reloaded in for each run. The county was setup using the small area datasets from the Irish Central Statistics Office (CSO). Small areas are the smallest level of geography that the census data is aggregated for and are made up of between 50 and 200 homes. Small areas for each town are loaded into the model with a single patch representing a small area. School location datasets are used to find the small areas that schools are located in within the county. Each small area patch records the real world distance between the center of the small area and the center of every other small area in the model. Agents are added to the town based on the closest census data to the year being modelled. Census data is produced every five years with the two most recent datasets for the 2011 and 2016 census. If we were modelling an outbreak from 2012 the 2011 census data would be used. However, if the outbreak was in 2014, the 2016 census data would be used. The number of agents in the simulation will be the number of people living in the town. The following steps are used in populating the town with agents and are performed for each individual small area:

- Each household is assigned a type (single, couple, couple plus others, couple with children, couple with children plus others, single parent, single parent plus others or other) and a family id that links all agents in the same household.

- Adults are added into each household. One agent is added to households with types single, single parent and other. Two agents are added to the households with type couple.

- Adults in each household are assigned a sex and age based on a probability distribution determined from the CSO census age, sex tables for the relevant small area.

  - The age categories provided by the CSO are by year until 19 after which ages are reported in ranges of five years, for example ages 20-24 or ages 60-64, and then anyone over 85 is combined into one age bracket. To have all ages represented in our model in each age bracket we randomly assign individuals one of the five ages represented in that bracket.

  - Couples are assigned opposite genders and ages within 10 years of each other.

- If a household type includes children a probability distribution determined from the relevant census data is used to determine if all children in the house are under 15, over 15, or both over and under 15.

- Data from the family units with children by size and age of children table is

used to determine the probability that each household with type child has 1, 2, 3, 4 or 5 children in the household.

- Children are added into each household.
- Children are assigned a sex and age based on a probability distribution extracted from the relevant census data and the type of children the household is assigned to (under 15, over 15 or both under and over 15).
- If the total number of agents populating a small area is not equal to the total number of agents who should be in the area based on the CSO data, additional agents are added and randomly assigned to households of types couple plus others, couple with children plus others, single parent plus others or other.
- All agents are assigned an economic status based on CSO data.
  - Agents over 65 are assigned to retired.
  - Agents between the ages of 5 and 14 are assigned to student.
  - Agents between the ages of 15 and 18 are first assigned to student. If there are more agents aged 15 to 18 in the small area than the number of students in the same age categories then the agents are assigned to looking for first job. If there are still more agents aged 15 to 18 they are then assigned an economic status of work, unemployed or sick/ disabled following the distribution for these categories for the relevant small area.

- Adult agents under 65 are assigned to work, looking for first job, unemployed, sick/ disabled or stay at home following the distribution for these categories for the relevant small area. Agents are only assigned to stay at home if they are part of a couple.
- Agents under the age of 5 are assigned to student if they have no stay at home parent. If they have a stay at home parent then a probability determines if the agent will be assigned to student.

- Agents with an economic status of work are assigned to a workplace based on commuting data. The data is used to determine the probability that an agent in a given electoral division (CSO area one step above small areas) will commute to any other electoral division. The commuting data also includes agents who commute out of the region, if an agent ends up commuting out of the region they will select a patch in the model that is not designated as a small area and commute there during work hours.

- Agents aged 13 and up with an economic status of student are assigned to a secondary school based on a distance function. The closer the secondary school to the student the greater the probability they student will attend the school.

- Agents aged 4 to 12 with an economic status of student are assigned to a primary school based on a distance function. The closer the primary school to the student the greater the probability they student will attend the school.

- Agents aged 3 and below with an economic status of student are assigned to a pre-school closest to their home patch.

- When all agents in the region are assigned a workplace or school social networks are created.

  - Agents create a family network with all other agents in their household.

  - Agents create a work network with all other agents that work in the same workplace.

  - Students create a school network with all other agents in their school and a class network with other agents in the school who are the same age as them.

- If vaccinations are included in the model agents are given an immunity level based on the disease being modeled and vaccination data for that disease. Irish vaccination data is used to determine the percentage of each age group that have received vaccinations for the infectious disease being modelled. For example, if 90% of 1 year olds in Ireland had been given the MMR vaccination in 2011 and we are running a model for 2012, we give each agent in the model with an age of 2 a 90% chance of having been vaccinated. If an agent is vaccinated they are given a 97% chance of being immune to the disease. This takes into account vaccination failure and is based on the vaccine effectiveness rate for measles (Nelson & Williams, 2007). Half of the agents with age less than 1 are given immunity to a disease to mimic passive

immunity infants receive from their mothers (Nicoara et al., 1999). For any agents that have an age corresponding to a vaccination year not in our data we give a 99% chance of being immune. Prior to vaccination campaigns the majority of the population would have either had or been exposed to childhood diseases, such as measles, leaving them immune in later life. If socioeconomic status is used in the model, the percent of individuals receiving vaccinations is adjusted based on the household socioeconomic status of the individual.

- Finally, a given number of agents are given the status of infected.

Input Data

The model does not use input data to represent time-varying processes.

Submodels

- **Clock:** The clock submodel keeps track of the time, day and week of the model. The time is the determined as the modulus of ticks and 12. When the time goes back to 0 the day is increased by one. If the day is 8, the week is increased by one and the day goes back to 1.

- **Move:** In the model agents move in one step between their current location and their desired destination. Some movements are predetermined by the models rule others are determined using a gravity model. A gravity model is used to determine agents movements throughout the region. This is an

alternative to the simple random movements used in earlier models. Random selection of the agents next destination is a naive model that was acceptable under the assumptions of a small town or city but when the area over which agents can travel increases to say a county this assumption of random movement does not hold as well. For example, in a small town moving to the other side of the town is much more likely than moving from one side of the county to the other. To account for this a gravity model is used to determine movements. Gravity models are a type of transportation model that is similar in formula to Newton's gravitation model. A traditional gravity model gives the interactions between two location pairs and determines those interactions based on the characteristics of a location and the distance between locations (Rodrigue et al., 2006). In the model, agents move between home and school or work at certain predetermined times and will return home at predetermined times. On weekends, summers for students and after school or work hours agents will move through the community these movements are determined by the gravity model. The probability of an agent moving to another small area is proportional to the population density of the small area, an area with more agents is more attractive, and inversely proportional to the distance to the small area from the agents current location, areas that are farther away are less attractive. We feel that this transportation model provides a more accurate model of movement within a larger area than that in the original town model.

The set predetermined movements are as follows:

- Agents who are working leave their home on the fourth time step of the day, which would be equivalent to between 8am and 9am, arrive at work over one time step, spend 4 time steps (8 hours) at work and then return home.
- Students also leave on the fourth time step but only spend 3 time steps (6 hours) at school.
- Stay at home agents who have children in primary school travel with their children to school on the 4th time step and then return home during the same time step. Between the fourth and seventh time step (when students return home from school) stay at home agents move throughout the town: at each step if an agent is at home they have a 50% chance of staying at home. If not at home, an agent has a 50% chance of picking a new destination in town with the gravity model and moving there. At the 7th time step of the day the stay at home agents will go to their child's school and then travel home.
- Stay at home agents who do not have a child in primary school move throughout the town using the gravity model between the fourth time step and the seventh time step the same way stay at home parents move when they are not travelling with their children to school or home.
- Agents younger than 4 who are not assigned to a preschool move with their stay at home parent throughout the day.

- Agents who are unemployed, looking for their first job, retired or sick/ disabled move throughout the town with the gravity model between the 4th and 10th time steps of the day.
- If an agent is infected with the disease simulated within a model then their behaviour is affected. Infected agents have a certain probability of staying home. If they are working, the agents will stay home 30% of the time. Students will stay home 70% of the time. Unemployed agents will stay home 75% of the time, and stay at home agents with primary school children will stay at home 10% of the time when accompanying children to school and 50% of the time when moving around town. Stay at home agents with non-primary school children will stay at home 50% of the time.
- All agents will move throughout the town on the weekends using the gravity model.
- After each move agents will make note of what other agents they have come into contact with using their social networks.
    * If an agent is at home they will come into contact with all other agents in their family network who are also at home.
    * If an agent is at school they will come into contact with other agents also in school who are in their class network a given percentage of the time and with other agents also in their school who are in their school network a given percentage of the time. (The user of the

model sets the percentage chance of contact but the chance that an agent comes into contact with another agent in their class network should be higher than the chance they come into contact with a member of their school network)

    * If an agent is at work they will come into contact with other agents in their work network a given percent of the time.

    * If an agent is in the community they will come into contact with other agents in the same small area also in the community a given percent of the time. This percent is determined by the social networks the agents are in. There can be a different chance for agents in the same family network, class network, school network, work network and agents who are not in each others social networks this is set by the user.

- **Infect:** When an infected agent comes into contact with a susceptible agent, the infected agent will determine if they will infect the susceptible agent based on the infection rate, a variable chosen at the start of the simulation. If the infected agent determines it will infect the susceptible agent, the susceptible agent will change their health status from susceptible to exposed.

- **Infect SEIR:** When the number of agents infected or exposed passes a certain threshold, that is set for by the user, the disease component of the model will switch from the agent-based to equation based and remain equation based until the number of agents infected or exposed falls below the

threshold again.

The equation based part of the disease component uses an SEIR difference equation model. Difference equations were chosen over the more common differential equation models because of the discrete time space that is used in difference equations. This is more analogous to the agent-based model and will allow for a more seamless transition between the two models. In the simulation, each geographic area selected runs its own SEIR difference equation model. The model can be run at the small area level, the town level or the county level. The equations are as follows:

$$S_{t+1} = S_t - \frac{\beta I_t S_t}{N} \tag{C.2}$$

$$E_{t+1} = E_t + \frac{\beta I_t S_t}{N} - \sigma E_t \tag{C.3}$$

$$I_{t+1} = I_t + \sigma E_t - \gamma I_t \tag{C.4}$$

$$R_{t+1} = R_t + \gamma I_t \tag{C.5}$$

Where S_t is the number of susceptible agents in the small area in the previous time step and S_{t+1} is the number of susceptible agents in the geographic area in the current time step. E_t and E_{t+1} are the number of exposed agents in

the small area in the previous and current time steps, I_t and $It + 1$ are the number of infected agents in the small area in the previous and current time steps, and R_t and R_{t+1} are the number of recovered agents in the small area in the previous and current time steps. β is the infection rate or the probability of infection per contact between agents, σ is the rate of moving from exposed to infected and γ is the recovery rate.

In a fully equation based disease component, each geographic area starts its difference equation model when an infected or exposed individual enters the area. In a hybrid model the difference equation model will start when the number of infected or exposed individuals is over a certain threshold. This could happen in two ways, either an agent from outside the area who is already exposed or infected moves into the area or an agent who is from the area becomes infected outside and returns home. Once the difference equation model has started it continues until there are no longer any more exposed or infected agents in the model.

At each time step, each area will calculate the values for the difference equations and adjust the number of agents in the area in each category. If the rounded difference between E_{t+1} and the number of exposed agents in the area is greater than 0, that number of susceptible agents in the small area will randomly be selected to move from the susceptible category to the exposed category. Similarly if the rounded difference between I_{t+1} and the count of infected agents in the area is greater than 0 than that number of exposed

agents will be randomly selected to move from exposed to infected. If the rounded difference between R_{t+1} and the count of recovered agents in the area is greater than 0, than that number of infected agents in the area will recover.

Because movement is possible, there are times when the total number of agents in the area in one of the four categories is less or greater than the value predicted in the model. Adjustment are made to account for this. If the value for E_t, I_t, or R_t is less than one and the count of agents exposed, infected or recovered in the area is greater than one then the value for E_t, I_t, or R_t is changed to the count of agents in that area who are exposed, infected or recovered. If the values for the difference between E_t, I_t, or R_t and the number of agents exposed, infected or recovered respectively in the geographic area is greater than the number of agents who could potentially move into the compartment (if the difference between E_t and the count of agents exposed is greater than the number of susceptible agents) the value for E_t, I_t, or R_t are adjusted down to reflect the actual counts of agents in the geographic area.

- **Recover:** An exposed agent will use a probability distribution to determine the number of time steps it will stay exposed before it becomes infected. Similarly, when the agent switches from exposed to infected the agent will use a probability distribution to determine the number of time steps before they are recovered/immune. Once an agent has recovered they cannot become

infected again.

- **Update Global Variables:** At the end of each time step, all global variables are updated. The counts and percent of susceptible, exposed, infected and recovered agents are all calculated. The average number of contacts across all agents in the model is calculated by taking the average of each agent's contacts.

- **Find Contacts:** If the user has selected to keep track of contacts, find contacts will run each each step. Each agent has a vector that has recorded the number of other agents they have come into contact with each time step. Find-contacts has each agent find the average total contacts and add this to a list of average contacts for all agents. The average across all agents is then taken to find the average number of contacts by an agent each time step.

- **Who-sick:** A sub model to find and report the ID, age and home location of the agents who became infected in a given time step.

Appendix D

List of Publications

Journal Articles

Hunter, E., Mac Namee, B., & Kelleher, J. (2018) An open-data-driven agent-based model to simulate infectious disease outbreaks. PLOS ONE https://doi.org/10.1371/journal.pone.0208775

Hunter, E. Mac Namee, B. & Kelleher, J. (2018) A Socioeconomic Segregation Model to Help Setup an Agent-Based Model for Infectious Diseases. Journal of Artificial Societies and Social Simulation. 21 (4) 9 http://jasss.soc.surrey.ac.uk/21/4/9.html

Hunter, E. Mac Namee, B. & Kelleher, J. (2017) Taxonomy for Agent-Based Models in Human Infectious Disease Epidemiology. Journal of Artificial Societies

and Social Simulation. 20 (3) 2.

http://jasss.soc.surrey.ac.uk/20/3/2.html

Conference Papers

Hunter, E., Mac Namee, B., & Kelleher, J. (2019) Degree Centrality and the Probability of an Infectious Disease Outbreak in Towns within a Region. ESM 2019 Conference.

Hunter, E., Mac Namee, B., & Kelleher, J. (2018) A Comparison of Agent-Based Models and Equation Based Models for Infectious Disease Epidemiology. AICS 2018 Conference.

Hunter, E., Mac Namee, B., & Kelleher, J. (2016) An Open Data Driven Epidemiological Agent-Based Model for Irish Towns. AICS 2016 Conference

Appendix E

List of Employability and Discipline Specific Skills Training

Employability Skills

- **GRSO 1001 Research Methods** 5 credits. This course was taken at the start of my PhD to refine my research skills.
- **SPEC 9160 Problem Solving, Innovation and Communications** 5 credits. I took this course to gain important communication skills as well as to learn new problem solving methods.
- **GRSO 1010 Introduction to Pedagogy** 5 credits. This course was selected to help gain skills for lecturing.
- **PRJM 2000 Project Management** 5 credits. I took this course to help get a better grasp on project management related to PhD research.

Discipline Specific Training Skills

- **UCD CSTAR Fundamentals of Epidemiology** 5 credits. This was an external module and part of the UCD CSTAR Summer School in Epidemiology. It was taken to improve my background in the field of epidemiology.
- **MATH 9974 Biomathematics** 10 credits. This course was take to help improve my equation based modelling skills.
- **SPEC 9270 Machine Learning** 5 credits. I took this course to expand my knowledge in the field of computer science.

Additional Courses

- 11th Summer School in Individual and Agent-Based Modelling
- Agent-Based Models for Public Health (EPIC Summer School at Columbia University)
- Introduction to GIS in Public Health (EPIC Summer School at Columbia University)
- Introduction to Machine Learning for Epidemiologists (EPIC Summer School at Columbia University)