Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Catalyst: Agents of change
Integration of compartment and agent-based models for use in infectious disease epidemiology

Presented by: Shannon K. Gallagher

Accepted by: Department of Statistics & Data Science

Readers:

William F. Eddy, Advisor

Joel Greenhouse

Howard Seltman

Samuel L. Ventura

Approved by the Committee on Graduate Degrees:

Richard Scheines, Dean Date

Carnegie Mellon University

Catalyst: Agents of change

Integration of compartment and agent-based models for use in infectious disease epidemiology

A Dissertation Submitted

in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy

in

Statistics

by

Shannon K. Gallagher

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

July 2019

In loving memory of Gramps

Acknowledgements

Thank you to everyone who contributed to this dissertation, whether it be through direct advising, insightful conversation, or encouragement. I could not have completed this dissertation without your support.

In particular, I would like to thank Bill Eddy for being my advisor for nearly six years. His insight, advice, and anecdotes have been invaluable. Special thank you to Joel Greenhouse who went above and beyond the duties of a committee member with careful comments, great ideas, and helpful feedback. A big thank you also to Sam Ventura for helping me through the highs and lows of this process with edits, critiques, schedules, and caffeine. Thank you also to Howard Seltman for his advice along the way.

I would like to thank the many helpful past and present faculty members of the CMU Statistics & Data Science Department including Chris Genovese, Alessandro Rinaldo, John Lehoczky, Steve Fienberg, Amelia Haviland, Peter Freeman, Robin Mejia, Aaditya Ramdas, Nynke Niezink, Andrew Thomas, Brian Junker, and Ryan Tibshirani. A very special thank you to Rebecca Nugent for helping me through many a difficult spot. I also would like to thank CMU alum Dean Follman for help in obtaining data. Additionally, I would like to thank my great professors at Carnegie Mellon University who inspired me and encouraged me these past nine years: Greggo Johnson and Tom Bohman and especially to Judy Holdener and John Mackey.

Thank you also to the incredibly capable and helpful staff in the Department of Statistics & Data Science including Beth Dongilli, Christopher Peter Makris, Mari Alice McShane, Jess Paschke, and Carl Skipper. A special thank you to Margie Smykla for helping me out over the past six years!

To Carolyn Shetter and Madison Kerr, thank you for encouraging me from the very beginning. You two inspired me to keep going. To my undergraduate friends Zach Branson, Erin Taylor, Carson Sestilli, Shaina Mitchell, and Taylor Caligaris thank you for helping me get through a challenging and rigorous four years.

To my colleagues and friends in the graduate program at CMU and elsewhere: thank you to Lindsay Kohorn, Jonathan Fintzi, Liz Lorenzi, Joe Pane, Jacqueline Liu, Andersen Chang, Francesca Matano, Jerzy Wieczorek, Michael C. Vespe, Sam Adhikari, Beau Dabbs, Christine Dabbs, Bret Vukoder, Mikaela Meyer, Maria Jajha, Theresa Gebert, Manjari Das, Alden Green, Xiao Hui Tai, Neil Spencer, Xiaoyi Yang, Collin Politsch, Yotam Hechtlinger, Kevin Lin, Natalie Klein, Jackie Mauro, Maria Cuellar, Nic Dalmasso, Ciaran Evans, and Benjamin LeRoy. A special thank you to Lee Richardson for all the work with MIDAS.

To Emily Ruppel, Robin Dunn, Patrice Daniel, Corina Ramirez, and Weronika Balewski thank you for all the support. To Adelaide Cole, thank you for being there through the years. To Abby Smith, thank you for being a delight and a light in my life. To Kate Sickler, Neri, and Libby, thank you so much!

Thank you Purvasha Chakravarti for all your support and encouragement over the years.

Thank you to Taylor Pospisil for the tech support, support in athletic endeavors, and all the lunches!

Brendan McVeigh, thank you! I would not have made it through the past five years without you. You are a wonderful friend and office mate.

Honorable mentions go to Amanda Luby and Kayla Frisoli.

Amanda, thank you for the past five years of friendship and for always being ready with a funny comment in the office. I am so glad we are able to finish our dissertations in the same month!

Kayla, thank you. I am grateful and appreciative for your unwavering support and friendship. Thank you for challenging me to improve in many facets of my life! There is no one I would rather have on my team.

Of course, I need to thank my family for putting up with me for the past five years (and then the previous 22). To Sherrie, thank you for joining our family as my sister-in-law and being a source of inspiration for me. To Sean, thank you for challenging me, helping me, supporting me, and being the greatest big brother a little sister could ever have. You have always inspired me to be the best I can be.

I need to thank my dad and mom Michael and Kathee Gallagher who have been with me and supported me every single step of the way. For reading with me as a child, for picking me up after school from the advanced math classes, for all the fun times, for all the life lessons, and for getting to know you as an adult, thank you. I love you so much.

Finally, to my grandfather Terence Keegan, Gramps. Thank you for instilling in me a love of learning and the stubbornness to persevere. I wish you could have been here. This one is for you.

Abstract

Two common classes of models for infectious disease epidemiology are compartment models (CM) and agent-based models (AM). Despite similarities being noted between the two classes, little has been written to explicitly connect the two classes of models. This dissertation improves upon the statistical inference within infectious disease models that belong to either the class of CM or AM. Specifically, we 1) theoeretically relate CMs and AMs under both a common but stringent assumption of epidemic models and under more general conditions, 2) develop and refine model selection methodology for models the susceptible-infectious (SI) or susceptible-infectious-recovered (SIR) framework of epidemic models, and 3) apply our theory and methodology to historic and modern outbreaks.

Chapters 2-3 examine the statistical relationship between CMs and AMs and study the essential features of each, which include homogeneity of individuals within groups and homogeneous interaction between susceptible and infectious individuals. We show that under a broad set of conditions, a CM has an equivalent AM and an AM an equivalent CM, in terms of the number of individuals in each state at a given time. We then discuss the importance of this theoretical relationship, especially with respect to the total number of states required to adequately model an epidemic, which in turn determines the amount of heterogeneity of interactions in our model.

Chapter 4 presents methodology for selection of the the total number of states required to adequately model an epidemic for common epidemic models within the SI and SIR-framework. One method quantifies the level of homogeneity of interaction among agents, and visual diagnostics are developed to assess the model fit to observed data.

Finally, Chapters 5-9, comprehensively examine the statistical relationship and model selection methodology in two case studies: an outbreak of measles in Hagelloch, Germany (1861-1862), and an outbreak of Ebola in Western District, Sierra Leone (2014-2015). For both case studies, we perform model selection to determine the best fit CM-AM pair. We then use the selected CM-AM pair to simulate scenarios of interest to policy makers. Examples include: the general infection reduction, isolation and quarantine, school closure, sensitivity to initial conditions, and the importance of the value and interpretation of the

effective population size. We find that the CM-AM pair is a useful tool to analyze past epidemics as well as plan for future epidemics.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

On October 30, 1861 a German thirteen year-old became stricken with symptoms of what was later confirmed to be measles, in a town consiting of approximately 600 individuals. By January 30, 1862, the 188th and final case was recorded. All but 13 children in the village were infected, and the outbreak resulted in 12 fatalities.

On April 30, 2019, an adult resident of Pittsburgh, PA was confirmed to have measles and was treated at a local hospital. In contrast to the German outbreak, the Pittsburgh outbreak was contained and resulted in no fatalities. Over 150 years after the German outbreak, measles remains a threat to the well-being of individuals, despite the existence of an effective vaccine. Fortunately, in addition to significant advances in modern medicine such as vaccines, statistical modelling has developed as an important tool for prevention of and intervention of the outbreak of infectious diseases.

Statistical infectious disease modeling broadly focuses on either 1) prediction or 2) inference about a disease. The first aspect attempts to predict when and where new instances of a disease will occur. The second aspect attempts to learn information about a disease such as the infection rate, recovery rate, person-person interaction structure, how different diseases compare to one another, and how environmental and demographic characteristics of a population influence the outbreak of the disease. Both prediction and inference allow decision makers to better allocate resources, alert policy makers and the public, and implement prevention routines, but only the second aspect allows us to to learn *why* and *how* a disease transmits through a population.

In this dissertation, we focus on improving the second aspect: statistical inference of infectious disease. Specifically, we improve inference by 1) theoretically relating the statistical properties of two classes of commonly used epidemic infectious disease models, 2) using the theory developed to improve parameter estimation and model selection within these two classes of models through novel visual diagnostics and statistical investigation of heterogeneity of disease transmission, and 3) applying these techniques to real

world data to examine hypothetical scenarios such as the implementation of interventions like isolation and quarantine.

The introduction proceeds as follows. In Section 1.1 we examine, in detail, related work of the two common model classes, comparisons of the two classes, hybrid models resulting from combining the two classes, and how our contribution adds to the collective knowledge and advancement of these models. In Section 1.2, we examine in more detail how parameter inference, model selection, and diagnostics are commonly performed within these model classes and how we aim to improve the three aspects. In Section 1.3, we discuss past and present case studies of measles and Ebola along with our approach to modeling these diseases using real world data. Finally, in Section 1.4, we summarize our contributions and where they appear in the rest of the dissertation.

1.1 Model classes: compartment models and agent-based models

In this dissertation, we study and synthesize compartment models (CM) and agent-based models (AM). In infectious disease epidemiology, the origins of CMs date back to the early 1900s whereas AMs have only recently gained traction in the past two decades due to advances in computing (Kermack and McKendrick, 1927; Epstein, 2007). Both classes of models enjoy a rich history in epidemiology for both the prediction and inference of infectious diseases including plague, measles, Ebola, and more.

As CMs and AMs are used to answer the same sorts of questions in infectious disease epidemiology, it makes sense to compare the two classes of models. At a high level, CMs are historically equation-based and depend on the assumption of homogeneous interaction of the population. On the other hand, AMs are simulation-based and begin to incorporate heterogeneous interaction of the population. However, that is not to say CMs completely lack the flexibility to incorporate heterogeneous interaction. At the same time, it is typically possible to determine the equations associated with AMs. As such, it becomes natural to question where the boundaries are between the two classes of models and whether the two classes can be leveraged to create hybrid models.

In the following sub-sections, we examine the brief history of CMs and AMs and their advancements in infectious disease modeling. At the end of this section, we discuss our contribution to the collective knowledge of these two classes of models.

1.1.1 Compartment models

CMs describe the transition of objects among discrete compartments over time. In infectious disease epidemiology, these compartments reside within the Susceptible-Infectious (SI) framework to describe how a disease spreads through a population. Perhaps the most well known CM is the SIR model, which stands

for susceptible, infectious, and recovered, respectively, which was introduced by Kermack and McKendrick (1927). Since then, more compartments have been added (or removed) to provide a wide class of models to describe the evolution of objects within the SI-framework. Many such examples are found in Daley et al. (2001).

Anderson and May (1992) identify two important assumptions in CMs: 1) homogeneity of the population and 2) the law of mass action. The first assumption is the idea that all objects in a particular state or compartment will behave in the same manner. The second is a property borrowed from chemistry which says that the mass of the product of reactants is proportional to the mass of the reactants, or in terms of infectious disease compartment models, the rate of change of individuals in a compartment at the next time step is proportional to the number of individuals in the compartment at the current time step. While the law of mass action is often used in AMs, and is seen in every model in this dissertation, the assumption of homogeneity is highly controversial and we examine this assumption in detail.

Compartment models within the SI framework can be as simple the SI model or can be made to be quite complex. For instance, the CM described by Pandey et al. (2014) has 26 compartments! Other common CMs include MSEIR, MSEIRS, SEIR, SEIRS, SIR, SIRS, SEI, SEIS, SI, and SIS, where M stands for passive infant immunity and E for exposed but not yet infectious (Hethcote, 2000). CMs have been used to model a plethora of diseases including plague, HIV, influenza, Ebola, and more (Kermack and McKendrick, 1927; Anderson and May, 1992; Mills et al., 2004; Althaus, 2014).

Stochastic versions of compartment models have also been studied as to better fit real world data. Some of the first stochastic versions arise from the Reed-Frost framework (Abbey, 1952), which assumed that the number of infected individuals in the next generation was distributed from a Binomial with a certain probability and the current amount of susceptibles. These became known as chain Binomials as they could be recursively computed. Becker (1981) generalized chain Binomials by allowing a flexible probability of transition between generations. The idea of the next step's number of infections being dependent only on the current state naturally lead to Markov models. These Markov models have been thoroughly examined (Jacquez and O'Neill, 1991; Allen and Burgin, 2000; Daley et al., 2001). Gani and Yakowitz (1995) describe how to create confidence interval bounds for deterministic approximations of random processes. Recent Bayesian approaches have also been attempted such as those described in Lekone and Finkenstädt (2006) and Fintzi et al. (2017). Researchers such as Figueredo et al. (2014) and Banos et al. (2015) use the Gillespie (1976) algorithm to create stochastic versions of common compartment models. The Gillespie algorithm is a form of Monte Carlo sampling that samples events at a random time τ in which an infectious (susceptible) agent has a chance to recover (or become infectious) in such a manner that the underlying CM average shape is maintained. These methods are especially useful in the context of epidemiology as monotonicity is respected in both the number of susceptibles and the number of recovered. For both methods, the magnitude of the error is closely related to the step size of the calculations, with smaller time intervals typically leading

to smaller error. In general, stochastic versions of CMs maintain the underlying shape of the deterministic CM but may vary wildly in variance or distribution.

Although CMs are aggregate models, in the sense that they only track the numbers of individuals in each state, compared to which state each individual is in, work has been done to incorporate spatial information. Coupled CMs are the idea of running a single, unique CM for each region but allowing for migration among regions. These models allow for more heterogeneity but also require fitting a large number of parameters. Examples of these include the coupled SIR model of Rvachev and Longini (1985) which allows for migration among 52 cities across the world and more recent examples of metapopulation, which are discussed more below.

The CMs found in Anderson et al. (1986); Colizza et al. (2006); Hooten et al. (2010); Zhou et al. (2019) study the effects of adding multiple forms of heterogeneity into the models. For example, Colizza et al. (2006) examine a large global CM with diffusion of disease for global epidemics and seek to "characterize the level of heterogeneity" within the model. Hooten et al. (2010) examine a SIRS model and incorporate spatial-temporal modeling for different regions of the US to study influenza. They conclude that temperature covariates are very important in examining the spread of the disease. Zhou et al. (2019) examine heterogeneity of interaction by partitioning a SIR model into different sub-compartments for each of the S, I, and R compartments to include age structure in order to examine the effects of measles vaccinations in India.

In fact, the possibilities for infectious disease modeling are nearly endless, as described in the comprehensive overview of epidemic modeling Hethcote (1994), entiled "A Thousand and One Epidemic Models." He notes that compartments and epidemic models in general need to be tailored to consider the aspects of "epidemiological compartment structure, incidence, distributions of waiting times in the compartments, demographic structure, and epidemiological-demographic interactions."

As a researcher varies each of the aspects mentioned by Hethcote, more and more heterogeneity is incorporated into the CM. However, it may become to difficult to keep track of the models, in terms of equations, as the models become more and more complex.

1.1.2 Agent-based models

In response to the demand for increasingly complex and heterogeneous epidemic models, agent-based models (AM) were developed. Falling under the broader class of "simulations," AMs are used to simulate autonomous agents and their interactions within a constrained environment over time and are described as a "generative" mode of science (Epstein, 2007).

Two of the first AMs date back to the 1970s with Conway's Game of Life as described in Adamatzky (2010) and the segregation of communities of Schelling (1971). These AMs, upon inspection, are quite similar, and contain all the important aspects of what we would expect to find in an AM. In both these

models, the environment is partitioned into a lattice and agents occupy cells within this lattice. In Conway's Game of Life, an agent may either have a value of dead or alive and in Schelling's segregation model, agents are either one of two races or a "null" state. In both these models, an agent's future state is determined by its present state along with the present state of the other agents, in particular, their direct neighbors. This is known as "cellular automata." The major difference in these two models is that of deterministic versus stochastic interactions of the agents. In Conway's Game of Life, the states of agents in an AM are completely determined by their initial states. On the other hand, Schelling's model incorporates a stochastic process, where agents move to another state based on a (literal) flip of a coin. Because of this, the concept of running multiple instances of a particular AM with given initial parameters is important, as different random draws produce different results. Through this stochastic process, variability is introduced into the model.

As computers became more powerful and more accessible, AMs became an option as a "new kind of science" (Wolfram, 2002), neither an inductive nor deductive mode. The AM, Transportation Analysis Simulation System (TRANSIMS) from Los Alamos National Laboratory is a foundational work in this field. TRANSIMS is the first, large-scale, *data-driven* AM of its kind, meaning the agents are based on actual U.S. citizens from data from the U.S. Census including demographic characteristics such as race and age. Additionally, the agents include activity information such as commute time and occupation type. The goal of TRANSIMS is to examine the "transportation infrastructure effect on the quality of life, productivity, and economy" (Smith et al., 1995).

TRANSIMS has agents with both individual and household characteristics; environments with roads, workplaces, and households; and activity assignments which have been assigned probabilistically to the agents and activities through a "route planner." Smith et al. (1995) note that all models within TRANSIMS are probabilistic, but the program overall takes more of a results-oriented approach rather than examining the variation within the model. TRANSIMS builds on the celluar automata framework by dividing a region into a grid to have a large number of agents evolve in a (relatively) small amount of computational time. TRANSIMS is still in use and is available today. Moreover, its influence can be found in its successors such as MATSims and EpiSims (Waraich et al., 2009; Eubank et al., 2004), the former which continues the goal of examining traffic patterns whereas the latter examines the spread of disease with an AM framework.

More recently, AMs have been used to model the spread of infectious disease (Longini et al., 2004; Grefenstette et al., 2013). In this field, AMs are sometimes called Individual Level Models (ILM) or Individual Based Models (IBM), as the term "agent" is more commonly used to describe a biological pathagen as opposed to the individual who receives the disease. AMs in this field have been used for prediction, inference, and study of hypothetical prevention strategies (Eubank et al., 2010; Bajardi et al., 2011; Barrett et al., 2013; Liu et al., 2015a; Wang et al., 2016). Typically in these models, agents are assumed to be non-random, as are the environments, with the only random variation arising through transference of a disease through

activities of agents. Variance of estimates are reported through simulation results accumulated by running the model hundreds of times.

A popular representation for AMs is that of a network or graph-based framework. In this framework, the agent states (e.g. susceptible, infectious, recovered) are node colorings or labels and the directed edges are conditional probabilities of evolution of states. The graph then updates at each time step based on current states and edge weights. However, the graph-based approach is not exclusive to AMs as CMs are often described in this manner.

Some researchers closely utilize the structure of the graphs. For instance, Liu et al. (2015a) examine the property of "hubs," those individuals with many contacts, and examine whether vaccinating these hubs alone is enough to curb the full effect of an outbreak of a disease. Scheffer et al. (1995) examine the concept of "super individuals" which simply represent multiple agents of a certain group or class. In this way, Scheffer et al. can drastically reduce the number of nodes in the graph and correspondingly speed up computational performance. However, the details of condensing agents into a similar group have not been thoroughly examined from a statistical perspective. In Siettos et al. (2015), the researchers create an AM in Western Africa and use small world transitions and vary a parameter which controls the density of the connections in the graph.

Cressie et al. (2009) note that one of the most difficult issues with AMs is incorporating and keeping track of uncertainty in the model. Uncertainty in AMs has many sources, including sampling design, model specification, parameter settings, and initial and boundary conditions. Their recommendation is to use a hiearchical model to keep track of the different sources of uncertainty, beginning with data, then the process, and finally the parameters. They also recommend that the models be cross-validated. Furthery, they note that often AMs suffer from the problem of identifiability.

Although AMs have been used widely in fields such as ecology, sociology, epidemiology and more, their statistical properties remain largely unstudied. The most important work done with AMs with regards to statistics is found in Hooten and Wikle (2010), but the work focuses on modeling the underlying probability of evolving from one state to another rather than the statistical properties of the AM.

As the agents themselves are often estimated or imputed from other sources of data, uncertainty within the agents themselves is a topic of recent interest. See, for example, Barrett et al. (2008); Chao et al. (2010); Gallagher et al. (2018). The challenges of incorporating different sources of data are explored both by Cressie et al. (2009) and Gallagher et al. (2018). The latter provides detail on the data harmonization process and introduces an open-source `R` package in order to improve transparency and reproducibility in the agent generation process. Abar et al. (2017) provide a summary of available AM software tools available.

The primary shortcomings of AMs are two-fold: 1) aligning the model to reality and 2) having sufficient computational memory and time. Wallentin and Neuwirth (2017) describe this as the computational-predictive trade-off, and each of the AMs presented above use varied approaches to align models to reality

while maintaining acceptable computational performance. We discuss their issues more in Section 1.2. Wolkewitz et al. (2008) aptly summarizes the problems of AMs and epidemic modeling in general when they state to make models "as simple as possible but not simpler."

1.1.3 Comparing and combining CMs and AMs

Similarities between CMs and AMs have been noted by many researchers but relatively few papers have been written about these comparisons. Axtell et al. (1996) write that AMs must be aligned or "docked" to their underlying model, often empirically, so the two approaches may be compared. Rahmandad and Sterman (2008) compare deterministic CMs and their AM equivalents, specifically that of the SEIR model. They find that using a fully connected network of agents, results of the two were quite similar although not exact. Other network structures such as small world and ring lattice produce markedly different results. Additionally, Rahmandad and Sterman find that population size has little effect on their results.

Figueredo et al. (2014) compare established AMs with their stochastic-version CMs, produced by the Gillespie method. They compare the two methods in three case studies relating to cancer by fitting mixed effect models and comparing the results. They find that although the two models may look similar, they result in different distributions.

The conclusion from these studies, in general, is that CMs and AMs often produce similar results, but AMs may produce extra results due to being able to track individuals throughout time. Many studies reveal that AMs and CMs sometimes act the same and sometimes differently. For example, Yang et al. (2015) note that heterogeneous networks are not strictly more infectious than homogeneous networks and look for critical levels of infection. They note that it is the heterogeneity of infection risk and not heterogeneity of agent interaction that determines the likelihood of outbreaks. Moreover, even though researchers seem to value variability in their simulations, they typically only analyze the mean (Edwards et al., 2003; Chen et al., 2004; Vincenot et al., 2011).

Some modelers attempt to leverage the advantages of both CMs and AMs by combining them into hybrid models. Analyzing global versus local effects, Fahse et al. (1998) decompose the system into two different time scales where one feature evolves more rapidly than the second. From this, they are able to extract global parameters from the AM. Also in ecology, Wallentin and Neuwirth (2017) examine switching between equation-based models and AMs in a predator-prey model in order to examine the computational-predictive trade-off. The conclusion is that they obtain different results from different models but that AMs can indeed be useful in terms of computational and predictive performance.

Bobashev et al. (2007) create a hybrid model, based on the SEIR model. Their model uses homogeneous agents to better demonstrate the relationship between CMs and AMs. This hybrid model utilizes an AM when the number of infected individuals is below a pre-selected threshold and then switches to a CM when

the number of infected is large. Their idea is that when the number of infected is large enough, the outbreak is stable enough to model through CMs, an idea also related by Jaffry and Treur (2008). This threshold is heuristically determined. The intuition is that heterogeneous effects are most important at the beginning and end of an outbreak and hence need a more detailed model at those times.

Banos et al. (2015) create a hybrid model, which they describe as a metapopulation model, that uses a SIR model within cities and agents traveling between them. Hanski (1998) describes metapopulation as the technique of reducing individuals and their ecosystem into a network structure. In this way, individuals grouped in a similar environment are assumed to behave the same way, yet migration is often allowed among the sub-populations. Banos et al. (2015) compare this hybrid model to a coupled SIR model in cities with instantaneous travel of agents and find that although results are similar when looking at aggregate totals of individuals, the models diverge when prevention strategies such as quarantine and avoidance are applied. The idea of metapopulation hybrid models is also explored in Bajardi et al. (2011), which studies the spread of the 2009 H1N1 pandemic. Here, a SEIR-like model is implemented within countries and individuals are able to travel among them, thus allowing them to analyze prevention strategies such as travel bans.

Another meta-population model is examined in Bradhurst et al. (2015) which studies the spread of foot and mouth disease in cattle. In this model, instead of CMs and equations controlling the diffusion of disease among communities, the communities themselves are treated as agents that interact with one another, and instead, CMs are used to track the spread of disease within communities.

The idea of hybrid models is generally well-received but the details, statistical and otherwise can be sparse and situation-dependent. Our contribution to the improvement of the collective knowledge of CMs, AMs, and any hybrids is showing statistically how the two classes of models are equivalent under certain (often light) conditions. This in turn can be used to help standardize parameter estimation and model selection.

1.2 Parameter estimation and model selection

All the models presented above rely on paremeters such as wait times between states (e.g. infection rate (β) and recovery rate (γ)) or number of individuals in a population (N) to determine results of the epidemic. *How* these parameters are chosen or estimated varies from model to model. Moreover, the problem of model selection is a difficult problem in itself, even when comparing models within the same class (e.g. a SIR model among the entire population vs. a SIR model split between children and adults).

Common methods for parameter estimation methods include minimizing an objective function (e.g. mean square error (MSE), mean absolute error (MAE), or squared norm of probability), regardless of distributional assumptions of noise (Brooks et al., 2015; Nakamura et al., 2017). Another method for parameter estimation method that is commonly used is maximum likelihood, which can be either parametric or non-parametric, although parametric models seem to be more common due to their interpretability (Cressie et al., 2009;

Shrestha et al., 2011; King et al., 2015; Venkatramanan et al., 2018). These likelihood based methods can be interpreted in either a frequentist or Bayesian method depending on assumptions about the data and the parameters (Wheeler and Waller, 2008; He et al., 2009; Venkatramanan et al., 2018).

Parameters commonly estimated in CMs and AMs include β the infection rate; γ, the recovery rate; $\mathcal{R}_0$ (average number of new infections when an infectious individual is introduced to a completely susceptible population), the reproduction number; ω, the serial interval (time between primary and secondary infection), the peak infectious percent, the peak day of infection, the final size of an epidemic (total amount of the population which has been infected over the course of an outbreak), and the outbreak duration. In addition to parameter point estimates, parameter variation and CIs also must be estimated. Gallagher et al. (2019) survey nine ways $\mathcal{R}_0$ and its sampling error are estimated and demonstrate the differences with an application to the 2009 influenza pandemic in the US.

Along with parameter variation estimation, sensitivity analysis of parameters is commonly performed in both CMs and AMs. Sensitivity analysis allows researchers to see how robust model results are to their parameter selections or estimates. The difference between sensitivity analysis and parameter variation estimation is that sensitivity analysis is typically not associated with distributional assumptions of noise whereas variation estimation is. Examples of sensitivity analysis are seen in Rahmandad and Sterman (2008); Chao et al. (2010); Hunter et al. (2018).

Model selection is also important in epidemic modeling. The first step of model selection is typically informed by medical experts and indicates which disease-level states individuals in a population may occupy in an epidemic. Common states are susceptible (S), exposed (E), infectious (I), and recovered (R) but other states include immune (M), funeral (F) transmission, and hospital (H) transmission (Venkatramanan et al., 2018).

Once the disease-level states are determined, model selection still must be done with the class of models restricted to the disease-level states. Common ways to select models are cross-validation, or minimizing an objective function, perhaps with a penalty on the number of parameters, such as in the Aikaike Information Criterion (AIC) (Wasserman, 2004; Cressie et al., 2009). For infectious disease modeling, cross-validation may be particularly difficult as data are often only seen for one epidemic "cycle" and the outbreaks themselves are vary both temporally and spatially. Cross validation has successfully been performed for cyclical and seasonal diseases such as influenza using a technique known as leave-one-season-out cross validation (Brooks et al., 2015).

Another aspect to consider in model selection is the level of homogeneous or heterogeneous interaction of indviduals within the model. Colizza et al. (2006) address this issue by measuring the level of heterogeneity of disease prevalence with entropy and using the entropy value in a hypothesis test for networks.

We improve model selection for the SIR-framework by introducing two novel visualizations, a log-linear plot and a ternary plot. The log-linear plot is derived from a recent theoretical result about SIR ordinary

differential equations (ODE) (see Harko et al. (2014)). This theoretical result allows us to transform the observed SIR data so that $\mathcal{R}_0$, the reproduction number, may be interpeted as the slope of the line through the transformed data. This, in conjunction with weighted linear regression, where the weights are estimated as the plug-in estimates of the inverse of our variance calculations of the number of individuals in each state, allow us to develop prediction intervals that empirically cover 95% of the observed data. As such, a plot of this nature allows modellers to assess whether the SIR-framework is a good fit for the data.

The second visualization, the ternary plot, is even more flexible than the log-linear plot, since it can be used to assess multiple groups within the SIR-framework (e.g. children and adults). Ternary plots are used to visualize SIS (which is comparable to SIR data) models in Safan et al. (2006), specifically to examine theoretical equilibria. Our diagnostic ternary plot extends and improves upon these theoretical SIR visualizations by introducing observed data to the ternary plot, adding 95% confidence regions, and visualizing the time scale. Our ternary plot can be used in a number of situations to assess the fit of a model.

To improve inference and to emphasize preference towards simpler models, we introduce a statistical investigation specific to the SI-framework to determine whether an agent interaction structure is homogeneous "enough" so that it can be treated as a simpler stochastic CM, which is associated with faster run times and possibly a more interpretable model. We also discuss the limitations of this statistical investigation.

1.3 Applications and decision making

Our final contribution to the improvement of inference in epidemic disease modeling for CMs and AMs involves two applications to real world data. We analyze two different scenarios in order to highlight different features of the CM-AM pairs.

The first scenario is an outbreak of measles in Hagelloch, Germany. The outbreak is quick and limited, occurring over a span of just over 90 days and involving fewer than 200 individuals. The Hagelloch outbreak is advantageous for testing epidemic models for two reasons: 1) because the outbreak is small and occurs in a fairly isolated village where the demographic population is fairly homogeneous, we are far more confident in making modeling assumptions about the population, especially in contrast to a large region with a heterogeneous population, and 2) the Hagelloch outbreak data contains rich demographic features and individual interaction structure such as shared households and school information.

The Hagelloch data was analyzed in Neal and Roberts (2004) and examined different population interactions such as household and classroom structure. Despite the data having originated in 1861, measles remains a threat to the well-being of indviduals all over the world, as evidenced by the recent outbreak in the USA (Stobbe, 2019).

Measles has been consistently studied in infectious disease epidemiology. For example, along with German measles (not to be confused with the outbreak of measles in Germany shown in this dissertation) and chicken pox, Abbey (1952) identifies measles as a disease that can be modeled, using the Reed-Frost Binomial transition model. Anderson and May (1992) note that measles is one of the most infectious diseases in terms of its reproduction number, $\mathcal{R}_0$. Bhadra et al. (2011) note that measles is a convenient disease to study due to its "clear clinical diagnosis, direct human-to-human transmission, lifelong immunity following infection, and the availability of extensive spatio-temporal incidence data." Chris et al. (2012) also analyze the Hagelloch data through the use of an SEIR model and He et al. (2009) analyze the outbreak of measles in a boys' dormitory in order to explain their "plug-and-play" likelihood based inference methods and software.

With regards to CMs and AMs, Zhou et al. (2019) examine SIR epidemic models with age structure to determine best vaccination strategies for measles. Liu et al. (2015a) studies the role of vaccination coverage for the control of measles outbreaks in California. Getz et al. (2016) also studies the role of vaccinations in measles outbreaks, this time using an AM. Hunter et al. (2018) use an AM to examine the spread of measles with regards to population dynamics in Ireland in 2012.

We improve upon existing measles analysis by examining the Hagelloch data. More specifically, we utilize the theory and model selection methods developed in this dissertation to first find an adequate CM-AM pair for the model. Once our initial model is found, we use the properties of the CM-AM pair to analyze hypothetical scenarios including general reduction of the infectivity of the disease, isolation and quarantine procedures, and school closing.

The second case study examines the recent Ebola outbreak in Western Africa, specifically in Freetown, Sierra Leone. The data covers a time span from 2014-2015 where over 8,000 individuals were infected in a population of approximately 1.4 million people. This Ebola outbreak is much larger in scale than the Hagelloch outbreak and allows us to show how CM-AMs can scale when the number of individuals in a population or the number of infections is large. However, despite the outbreak being larger in scale, the data for the Ebola outbreak we have access to contains little demographic information about the infected cases.

Ebola has been of recent interest in epidemic modeling due to outbreaks in Africa over the past decades. Ebola is also a disease of interest because of its unique transmission structure, as transmission can be passed on from those already deceased as well as transmitted through eating "bushmeat" (Rizkalla et al., 2007; Pandey et al., 2014; Siettos et al., 2015). Studies have been focused on the areas of the Democratic Republic of the Congo, Guinea, Liberia, and Sierra Leone (Rizkalla et al., 2007; Althaus, 2015; Siettos et al., 2015; Ajelli et al., 2016; Nakamura et al., 2017). Both CMs and AMs have been used to estimate $\mathcal{R}_0$, estimate disease parameters in the presence of complex interaction structures such as households, hospitals, Ebola treatment units (ETU), and contact tracing (Ajelli et al., 2016; Brown et al., 2016).

Special focus has been concentrated on decision making for regions infected with Ebola. For example, "ring trials" focus on vaccinating the contacts and contacts of contacts of the primary infected cases (Henao-Restrepo et al., 2017). Backer and Wallinga (2016) represent the outbreak as "network of local epidemics...to effectively control emerging infections."

In our study of Ebola, we examine high-level concepts such as spatial spread due to heterogeneous individual interactions, sensitivity to initial infection locations, and the importance of the value of the effective population size, N. The first two concepts have been studied in Henao-Restrepo et al. (2017) and Shrestha et al. (2011), for instance. We extend these studies through special emphasis on the variance of our estimates which include peak infectious day, peak infectious percentage, epidemic duration, and final size. For the third aspect, the effective population size (not to be confused with time-varying population size), few if any prior analyses are available.

1.4 Dissertation structure

To summarize, the goal of this dissertation is to improve upon current inference in statistical disease modeling for CMs and AMs. We improve inference through the three following steps:

1. Theoretical contributions linking the similarities between CMs and AMs
2. Improvements in model selection via methodology for quantifying heterogeneity and visual diagnostics
3. Demonstrations of theory and methodology via applications to real world data.

The structure of this dissertation is given in the following manner. In Chapter 1 we have introduced the problem, described work by other researchers, and foreshadowed our contributions.

Chapters 2-3 address the first step of theoretical contributions for CMs and AMs. In Chapter 2, we introduce the problem using a concrete example of a CM within the SIR-framework. We show the expected value and variance of the number of individuals in each of the S, I, and R states at a given time in our Binomial transition model. Following that, we show how the CM may also be written in the form of an AM and be equivalent in distribution in terms of the number of individuals in each state at a given time. We then show how a CM with a specific (but useful) form may be written as an AM and call this the CM-AM pair.

In Chapter 3, we relax the assumptions on the form our CM may take and show that under more general assumptions, each CM has an AM pair. We demonstrate this concept with an example which we call the "lock-step" model. Additionally, we show that any AM has an equivalent CM, but this CM-AM pair is dependent on the total number of states allowed in the CM. At the end of the chapter, we discuss the implications of our results and how they guide model selection.

Chapter 4 addresses the second step of improving parameter inference and model selection, limited to the SI and SIR-frameworks. We introduce two novel visualizations which can aid researchers in assessing model selection for models within the SIR-framework. Additionally, we introduce a statistical investigation to determine whether an agent interaction structure is homogeneous "enough" in that the epidemic model can be adequately modeled through a simpler CM with homogeneous interaction of agents.

Chapters 5-9 comprehensively demonstrate the theoretical and methodological contributions presented in Chapters 2-4 using two real world data applications. Chapters 5-7 examine an outbreak of measles in a small town in Germany during the 1860s. In Chapter 5, we examine the data and find a best fit CM-AM pair. After we find our best fit model, we examine hypothetical scenarios with our AM. In Chapter 6, we examine how the results of the outbreak may have changed if we were able to generally reduce the infectivity of the disease across the entire population. Then in Chapter 7, we examine more tangible intervention scenarios such as isolation and quarantine of infectious agents and closure of the school.

In Chapters 8-9, we study the Ebola outbreak of 2014-2015 in Western District, Sierra Leone, which includes the capital of Freetown. In Chapter 8 we examine the Ebola data set and perform model selection within the SIR-framework. We find N, the effective population size to be particularly important in selecting the best model within this framework. After model selection, in Chapter 9, we study hypothetical scenarios with our selected CM-AM pair. Specifically, we perform experiments to examine the effect of heterogeneous versus homogeneous interaction of agents, sensitivity of the model to the location of initial infectors, and the size of the effective population.

Finally in Chapter 10 we summarize the results of the dissertation and discuss directions for future work.

Chapter 2

Statistically relating CMs to AMs

2.1 Background

Agent-based models (AM) have become an increasingly more common model class used to study the spread of disease throughout a population over a period of time. For example, recent use cases of AMs include developing mitigation strategies for the spread of smallpox, predicting the effect school closures have on the spread of influenza, and inferring the infectiousness of measles given a complex network structure (Eubank et al., 2004; Grefenstette et al., 2013; Liu et al., 2015b).

Recall, that AMs, as described in Chapter 1, model the movement of individuals, known as agents, through states (e.g. susceptible, infectious recovered) over time, where the movement of the agent is dependent on the states of the other agents as well as the environment.

Like any other statistical model, the validity of inference from an AM depends on the assumptions underlying the model. Specifically, these assumptions include validity of the agents themselves, validity of the environment in which the agents interact, validity of the interaction structure among the agents and their environment, and validity of disease parameters used in the simulations. If these assumptions hold, we can trust inferences from our AMs which, in turn, can help researchers explore unique hypothetical situations. See for example Carley et al. (2006) which explores reactions to a hypothetical release of anthrax in Washington D.C. Checking the validity of these assumptions is known in the AM literature as "docking" (Axtell et al., 1996).

We address the problem of docking by relating AMs theoretically to compartment models (CM), which have been used in epidemiological modeling to predict and study the outbreak of diseases for nearly one hundred years (Anderson and May, 1992). For the subject of epidemiological modeling, CMs have been used from a statistical perspective for over 70 years (Abbey, 1952) and many papers and books have been written about the properties of CMs. In particular, numerous papers and books have been written about how to fit

statistical disease models to data. In other words, statisticians and epidemiologists have spent much time and effort to dock CMs to reality. For example, recent software packages developed to aid the researcher in fitting CMs to data include `tSIR` and `pomp` (King et al., 2015; Becker and Grenfell, 2017), which are both available for the `R` language. The prevailing idea is that improved docking of AMs, will, in turn, lead to better inference with AMs, which will ultimately allow for the better understanding of disease outbreaks. If we can relate AMs to CMs in a statistically useful manner, then we have the whole literature of model fitting available to us to improve inference in AMs.

In this chapter, we integrate AMs into the established statistical framework for CMs. We do this by introducing (or incorporating) standard statistical terminology for deterministic epidemiological models, stochastic CMs, and stochastic AMs. Following that, we show that given a fixed deterministic model we can create equivalent CM-AM pairs in terms of equal number of individuals within different states (e.g. susceptible, infectious, or recovered). Finally, we show that a fixed deterministic model is not required and that we can create an equivalent AM for any existing CM and *vice versa*. This equivalency is based on the same parameters used in both models and hence estimation of these parameters in one framework is equivalent to estimation of the parameters in the other.

This chapter proceeds as follows. In Section 2.2 we describe the CM-AM equivalence with a concrete example, the Kermack and McKendrick SIR. We introduce concepts such as the deterministic transition matrix, the stochastic CM, and the stochastic AM. We also show how a particular CM-AM pair is equivalent given the deterministic transition matrix for the SIR model. In Section 2.3, we generalize the concepts in Section 2.2 to a large class of models and show how given a deterministic transition matrix, we can have equivalent CM-AM pairs. Finally, in Section 2.4, we summarize the chapter.

2.2 A minimal example: Kermack and McKendrick SIR model

We begin with an example based on the Kermack and McKendrick (1927) deterministic SIR-CM, which we will refer to as the K&M model. The original K&M model is presented as continuous time differential equations. Unless otherwise noted, we will instead use the discrete time version (replacing d with Δ and with $\Delta t = 1$). We use the discrete time version because of three reasons: 1) data is always collected in discrete time; 2) the equations may always be re-scaled to any appropriate time unit; and 3) agent-based models are inherently discrete time models.

The K&M model describes how individuals move from the susceptible state (S), to the infectious state (I), and finally the recovered/removed state (R) over time given the initial number of individuals in each state. The model is commonly shown graphically in Figure 2.1.

The K&M model assumes the following properties:

1. Individuals may occupy only one state at a given time.

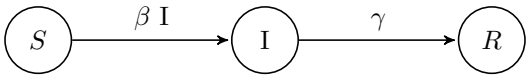

Figure 2.1: Graphical representation of the K&M SIR model. The circles represent states of individuals: Susceptible, Infectious, or Recovered, respectively. The arrows represent how individuals may move from one state to another over time. The expressions above the arrows represent the rates at which the individuals move from one state to the next.

2. Individuals within the same state at time t are homogeneous. That is, they are indistinguishable from one another.

3. All individuals within the total population mix homogeneously.

In this model, one or more infectious individuals are introduced to a population at time $t = 0$, giving a total population size of N. The susceptible individuals subsequently have the possibility to become infected and recover over time. The numbers of non-random susceptible, infectious, and recovered individuals at time t are $S(t), I(t), R(t)$, respectively. We assume the total population size is fixed and so

$$S(t) + I(t) + R(t) \equiv N.$$

The rate of individuals moving from one time step to the next (Δt) is given by Eq. (2.1), where β is the average rate of infection given in (time unit)$^{-1}$ and γ is the average rate of recovery given in (time unit)$^{-1}$.

$$\begin{cases} \frac{\Delta S}{\Delta t} &= -S \times \beta \frac{I}{N} \\ \frac{\Delta I}{\Delta t} &= S \times \beta \frac{I}{N} - I \times \gamma \\ \frac{\Delta R}{\Delta t} &= I \times \gamma \end{cases} \tag{2.1}$$

Since the total population is constant then it follows that

$$\frac{\Delta S}{\Delta t} + \frac{\Delta I}{\Delta t} + \frac{\Delta R}{\Delta t} \equiv 0,$$

and so one of the rates of change for one of the states is completely determined by the rates of change by the other two states. In this example, it is clear that $\frac{\Delta I}{\Delta t} = -\frac{\Delta S}{\Delta t} - \frac{\Delta R}{\Delta t}$.

Another way to write Eq. (2.1) is through deterministic transition matrix $\mathbf{D}(t)$, a $K \times K$ matrix where entry $D_{ij}(t)$ gives the non-negative number of individuals moving from state i to state j from time $t-1$ to

t. For the K&M model, $K = 3$ and $\mathbf{X}(t) = (S(t), I(t), R(t))^T$,

$$\mathbf{D}(t) = \mathbf{p}(t)\mathbf{X}(t) = \begin{pmatrix} 1 - \beta\frac{I(t)}{N} & \beta\frac{I(t)}{N} & 0 \\ 0 & 1-\gamma & \gamma \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} S(t) \\ I(t) \\ R(t) \end{pmatrix}$$

The difference equations in Eq. (2.1) are recovered if we take $\frac{\Delta \mathbf{X}}{\Delta t} = (\mathbf{D}(t)^T - \mathbf{D}(t))\mathbf{1}$, where $\mathbf{1}$ is a vector of ones. The form $\mathbf{D} = \mathbf{pX}$ incorporates the assumption "law of mass action," a property borrowed from chemistry which says that the mass of the reactants is proportional to the mass of the products (Anderson and May, 1992). Translating this to infectious disease epidemiology, we say that the number of individuals *moving out of* state i from time $t-1$ to t is proportional to $X_i(t-1)$, the number of individuals into state i at time $t-1$.

Notice that the change in the number of individuals in susceptible states is proportional to the old number of individuals in susceptible states. The difference in the number of individuals in susceptible states from time $t-1$ to t is dependent not only on β but also the percent of individuals who are infectious at time $t-1$, $I(t-1)/N$. In contrast, the difference in number of individuals in the infectious state from time $t-1$ to t is dependent only on the previous number of individuals in the infectious state and γ, the average recovery rate.

Originally, the K&M model was used to estimate the spread of plague in England. In particular, Kermack and McKendrick devised a method to efficiently *solve* for (as opposed to estimate) β and γ in the presence of data.

2.2.1 The stochastic SIR-CM

Since its debut in 1927, statisticians have converted the K&M model into stochastic counterparts which may account for noise. These include the Reed-Frost chain Binomial presented by Abbey (1952), numerous Markov processes described by Daley et al. (2001), and the "plug-and-play" methods described by He et al. (2009).

In particular, we will focus on an adaptation of the Reed-Frost chain Binomial with probabilities that mimic the original K&M equations. In particular, the stochastic SIR-CM we use is

$$\begin{aligned} Z_{t-1,S}|S_{t-1}, I_{t-1} &\sim \text{Binomial}\left(S_{t-1}, \beta\frac{I(t-1)}{N}\right) \\ Z_{t-1,R}|S_{t-1}, I_{t-1} &\sim \text{Binomial}\left(I_{t-1}, \gamma\right) \end{aligned} \tag{2.2}$$

$$
\begin{aligned}
S_t \,|S_{t-1}, I_{t-1} &= S_{t-1} - Z_{t-1,S} \\
I_t|S_{t-1}, I_{t-1} &= N - S_t - R_t \\
R_t \,|S_{t-1}, I_{t-1} &= R_{t-1} + Z_{t-1,R},
\end{aligned} \tag{2.3}
$$

with $(S_0, I_0, R_0) = (S(0), I(0), R(0))$.

In words, the new number of individuals in the susceptible state at time t, conditioned on the number of individuals in all states at time $t-1$, is equal to the old number of individuals in the susceptible state at time $t-1$ minus a Binomial draw with the size argument as the old number of individuals in the susceptible state and a probability based on both β and the proportion of infectious individuals at time $t-1$.

Similarly, the new number of individuals in the recovered state at time t, conditioned on the number of individuals in all states at time $t-1$ is equal to the old number of individuals in the recovered state at time $t-1$ plus a Binomial draw with the size argument as the number of individuals in the infectious state at time $t-1$ and the probability argument as γ.

Finally, in Eqs. (2.2)-(2.3) we set the initial values to those in the deterministic K&M model. That is, we assume the initial values are known.

We use the model presented in Eqs. (2.2)-(2.3) for a number of reasons: 1) Binomial draws conditioned on the previous number in states are an intuitive way to model an outbreak, 2) the S state is monotone non-increasing and the R state is monotone non-decreasing, just like in the K&M equations, and 3) the model is unbiased with respect to the original K&M equations.

We first determine the expected value and variance of the states in the models in Eq. (2.2).

Theorem 2.1. *Assume we are given a deterministic SIR-CM as in Eq.* (2.1) *and a corresponding stochastic SIR-CM as in Eq.* (2.2)-(2.3). *Then the expected value of the stochastic CM in terms of the state sizes at each time step is unbiased. That is, for all* $t = 0, \ldots, T$

$$
\begin{aligned}
E[S_t] &= S(t) \\
E[I_t] &= I(t) \\
E[R_t] &= R(t).
\end{aligned}
$$

In Theorem 2.1, the left hand side is the expected number of individuals in the susceptible, infectious, and recovered states, respectively from stochastic SIR-CM and the right hand side is the deterministic number of individuals in the susceptible, infectious, and recovered states, respectively from the deterministic SIR-CM. The result of this theorem is that, on average, draws from the stochastic SIR-CM will mimic the shape of of the curves from the deterministic SIR-CM.

Proof. We show the proof for $E[S_t]$. The other states follow the same idea. The base case is

$$\begin{aligned}
E[S_1] &= E\left[S_0 - \text{Binomial}\left(S_0, \frac{\beta I(0)}{N}\right)\right] \\
&= E\left[S(0) - \text{Binomial}\left(S(0), \frac{\beta I(0)}{N}\right)\right] \\
&= S(0) - S(0)\frac{\beta I(0)}{N} \\
&= S(1)
\end{aligned}$$

For the other states, this also holds. That is, $E[I_1] = I(1)$ and $E[R_1] = R(1)$. A recursive relation is used through use of the law of iterated expectation. Namely,

$$\begin{aligned}
E[S_t] &= E\left[E\left[S_t | S_{t-1}, I_{t-1}\right]\right] \text{ (law of iterated expectation)} \\
&= E\left[E\left[S_{t-1} - \text{Binomial}\left(S_{t-1}, \frac{\beta I(t-1)}{N}\right) | S_{t-1}, I_{t-1}\right]\right] \\
&= E\left[S_{t-1} - S_{t-1}\frac{\beta I(t-1)}{N}\right] \\
&= S(t-1) - S(t-1)\frac{\beta I(t-1)}{N} \\
&= S(t).
\end{aligned}$$

□

In this proof, note that the first argument in the Binomial is a random variable and the second argument $\beta I(t-1)/N$ is non-random. This proof, however, still holds true if we allow $I(t-1) = I_{t-1}$ since S_{t-1} and I_{t-1} are independent of one another given the initial states.

Moreover, we can compute the model variance of the stochastic SIR-CM in Eq. (2.2).

Theorem 2.2. *The model variance for Eq.* (2.2) *is given recursively,*

$$\begin{aligned}
V\left[S_t\right] &= S(t-1)(1-p_{t-1})p_{t-1} + (1-p_{t-1})^2 V\left[S_{t-1}\right] \\
V\left[I_t\right] &= V[S_t] + V[R_t] - 2 \cdot \text{Cov}[S_t, R_t] \\
V\left[R_t\right] &= I(t-1)\gamma(1-\gamma) + (1-\gamma)^2 V\left[R_{t-1}\right] + \gamma^2 V[S_{t-1}] \\
&\quad - 2\gamma(1-\gamma) \cdot \text{Cov}\left(S_{t-1}, R_{t-1}\right) \\
\text{Cov}\left[S_t, R_t\right] &= -\gamma(1-p_{t-1})V[S_{t-1}] + (1-\gamma)(1-p_{t-1})\text{Cov}[S_{t-1}, R_{t-1}].
\end{aligned} \tag{2.4}$$

As a result of Theorem 2.2, we see that the model is dynamically changing in variance for each state for an increasing t. This makes intuitive sense as the uncertainty accumulates over time. We demonstrate

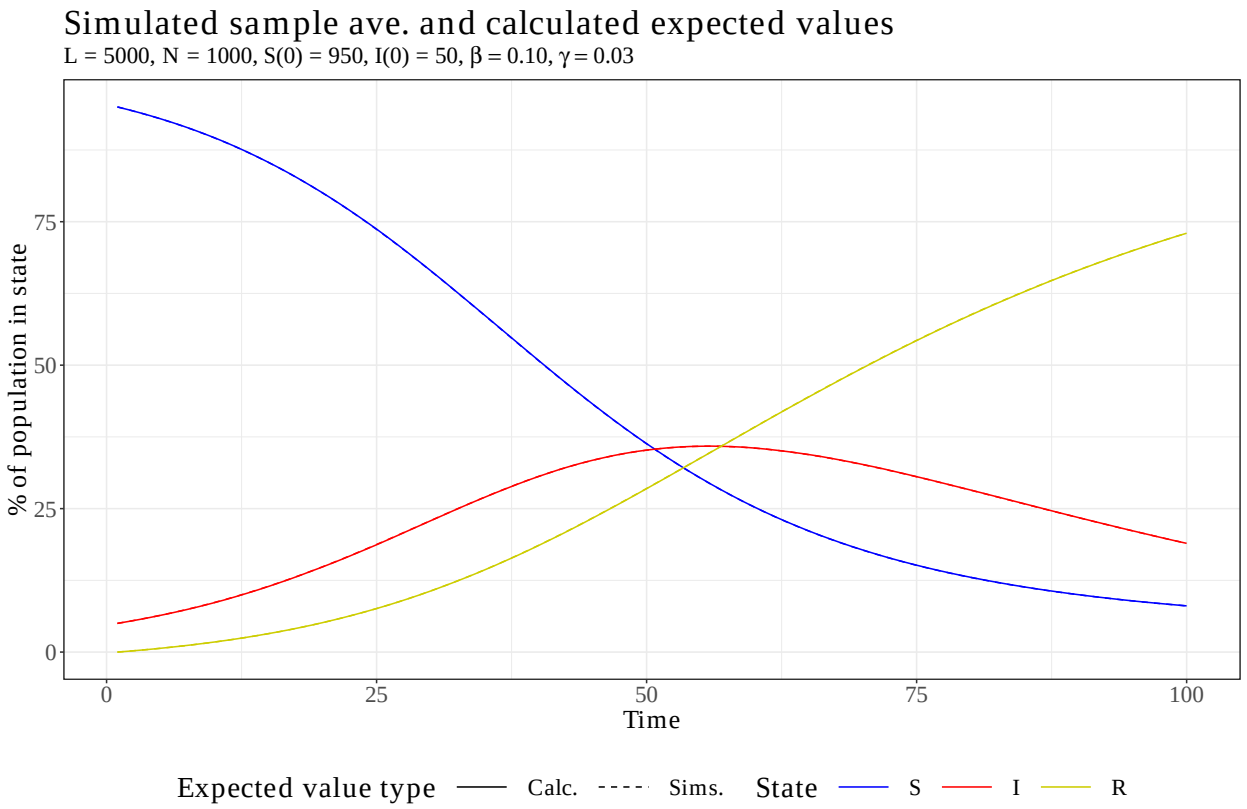

Figure 2.2: Expected value for each state for each time step from simulations (solid) and calculations (dashed). These sets of lines almost completely overlap. The simulations and calculations were generated with $L = 5000, N = 1000$, $S(0) = 950$, $I(0) = 50$, $\beta = 0.10$, and $\gamma = 0.03$.

this concept in the following simulations. We simulate an epidemic under the following settings: $N = 1000$, $S(0) = 950$, $I(0) = 50$, $\beta = 0.10$, and $\gamma = 0.03$. We generated $L = 5000$ sets of simulated data. The results are plotted in Figures 2.2 and 2.3. These figures show that the simulated expected values for each state and each time step match the calculated values as the sets of lines nearly overlap. Similarly, the sample variance from the simulations for each of the states at each time nearly overlaps with the calculated variance. Note that the peak variance for the infectious state $V[I_t]$ coincides with the peak of the infection.

2.2.2 The stochastic SIR-AM

We can also model the deterministic K&M SIR equations using an agent-based model (AM). We denote the state of the agent $n = 1, \ldots, T$ at time $t = 0, 1, \ldots, T$ as $A_{t,n} \in \{1, 2, 3\}$. Let $\{1, 2, 3\}$ represent the susceptible, infectious, and recovered states, respectively. Let the total number of agents at time t and in

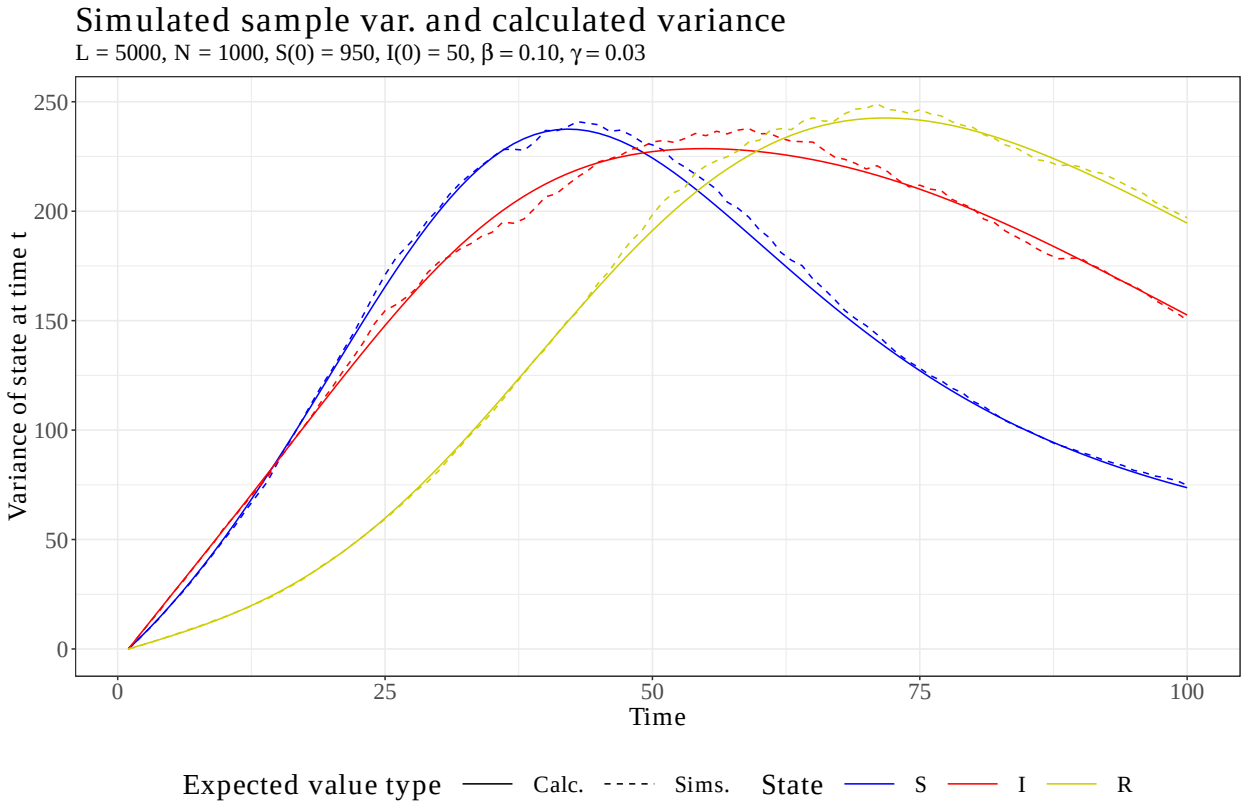

Figure 2.3: Variance for each state for each time step from simulations (solid) and calculations (dashed). These sets of lines almost completely overlap. The simulations and calculations were generated with $L = 5000, N = 1000$, $S(0) = 950$, $I(0) = 50$, $\beta = 0.10$, and $\gamma = 0.03$.

state k be given by the random variable

$$X_{t,k} = \sum_{n=1}^{N} \mathcal{I}\{A_{t,n} = k\}$$

where $\mathcal{I}\{\cdot\}$ is the indicator function of its argument. Finally, let $\mathbf{A}_t$ be the collection of all N agents at time t, $\mathbf{A}_t = (A_{t,1}, \ldots, A_{t,n})$.

Then the AM is specified by the following movement of the agents:

$$\begin{aligned} W_{t-1,n,S} &\sim \text{Bernoulli}\left(\beta \frac{I(t-1)}{N}\right) \\ W_{t-1,n,R} &\sim \text{Bernoulli}\left(\gamma\right) \\ A_{t,n}|\mathbf{A}_{t-1} &= \begin{cases} 1 + W_{t-1,n,S} & \text{if } A_{t-1,n} = 1 \\ 2 + W_{t-1,n,R} & \text{if } A_{t-1,n} = 2 \\ 3 & \text{if } A_{t-1,n} = 3 \end{cases} . \end{aligned} \tag{2.5}$$

In the stochastic K&M SIR in Eq. (2.5), the agents have a chance of moving to the next state, dependent on their current state, just like in the stochastic SIR-CM. However, now we distinguish between individuals, as each agent receives its own random update $W_{t-1,n,S}$ or $W_{t-1,n,R}$ depending on its state at time $t-1$.

Theorem 2.3. *Let there be a deterministic SIR-CM as given in Eq.* (2.1)*, a stochastic SIR-CM as given in Eq.* (2.2)*, and a stochastic SIR-AM as given in Eq.* (2.2)*. Then*

$$(S, I, R) \overset{d}{=} (X_1, X_2, X_3).$$

The take away from Theorem 2.3 is that the joint distribution for the number of states in the stochastic SIR-CM and the stochastic SIR-AM are equivalent. The proof of Theorem 2.3 is contained in a more general result, described in the following section.

2.3 General CM-AM pairs given a transition matrix

We can create equivalent CM-AM pairs for a much wider class of models than just the K&M deterministic SIR-CM. In fact, given any deterministic CM with transition matrix $\mathbf{D}(t)$ of size $K \times K$, we can create CM-AM pairs that are jointly equivalent in distribution in the number of individuals in each state $k = 1 \ldots, K$ and $\mathbf{X}_t$ is unbiased with respect to $\mathbf{D}(t)$.

Let $\mathbf{D}(t)$ be a $K \times K$ transition matrix for a deterministic CM, where K is the total number of states. Let entry $D_{ij}(t-1) \geq 0$ be the non-negative number of agents moving from state i to state j from time $t-1$ to t. Then $\sum_{j=1}^{K} D_{ij}(t-1) = X_i(t-1)$, the total number of individuals in state i at time $t-1$ and

$\sum_{i=1}^{K} D_{ij}(t-1) = X_j(t)$, the total number of individuals in state j at time t. We construct equivalent stochastic CM-AM pairs through the following process.

In the K&M SIR-CM, individuals only have one possible state to move to at a given time. Individuals within the susceptible state may only move to the infectious state, and infectious individuals may only move to the recovered state. Since there is only one choice (moving or not), we can use the fact that the sum of N independent and identically distributed (i.i.d) Bernoulli variables is equivalent in distribution to a Binomial variable drawing from size N,

$$\begin{aligned} W_n &\overset{iid}{\sim} \text{Bernoulli}\,(p) \\ Z &\sim \text{Binomial}\,(N,p) \\ \sum_{n=1}^{N} W_n &\overset{d}{=} Z. \end{aligned}$$

In more general systems, individuals may have more than one state they can move to, and we instead use the Multinomial/Multinomial equivalence shown in Eq. (2.6),

$$\begin{aligned} W_n &\overset{iid}{\sim} \text{Multinomial}(1,\mathbf{p}) \\ Z &\sim \text{Multinomial}(\text{N},\mathbf{p}) \\ \sum_{n=1}^{N} W_n &\overset{d}{=} Z, \end{aligned} \tag{2.6}$$

where $\mathbf{p} = (p_1, \ldots, p_K)$ is a vector such that p_j is the probability of moving from the current state to state $j = 1, \ldots, K$.

Then for a single state i, the difference equation is given by the sum of the number of individuals moving into state k and minus the sum of the number of individuals moving out of state k in a single time step,

$$\frac{\Delta X_i(t)}{\Delta(t)} = \left((\mathbf{D}^T(t-1) - \mathbf{D}(t-1))\mathbf{1}\right)_i. \tag{2.7}$$

Let $\mathbf{p}_i(t) = (p_{i1}(t), \ldots, p_{iK}(t))$ be a (non-random) vector of size K be the normalized transition rates, where

$$p_{ij}(t) = \frac{D_{ij}(t)}{X_i(t)}. \tag{2.8}$$

Stochastic CM for $\mathbf{D}_t$.

Assume there is a deterministic CM with states $\{k; k = 1, \ldots, K\}$, and assume the number of individuals in each state at each time step is given by the random variable $\mathbf{X}^{CM} = \left(X_{t1}^{CM}, \ldots, X_{tk}^{CM}\right)^T$. Further assume movements are given between pairs of states by deterministic, discrete time difference equations, a $K \times K$

matrix $\mathbf{D}(t)$, such that entry $D_{ij}(t) \geq 0$ where $D_{ij}(t-1)$ is the number of individuals moving from state i to state $j \neq i$ from time $t-1$ to t. Here, $p_{ij}(t)$ is the (non-random) probability of an individual moving from state i to j from time t to $t+1$.

The stochastic version of the CM relies on Multinomial draws $\mathbf{Z}_{ti}$ for $i = 1, 2, \ldots, K$ where $\mathbf{Z}_{ti} = (Z_{ti1}, \ldots, Z_{tiK})$ is a row vector,

$$\mathbf{Z}_{ti} \sim \text{Multinomial}\left(X_{ti}^{CM}, \mathbf{p}_i(t)\right).$$

That is, the number of individuals who move from state i to j from time t to $t+1$ is Z_{tij}, the jth entry of $\mathbf{Z}_{ti}$. The first argument is a random variable and the second is non-random. Define $\mathbf{Z}_t$ as a $K \times K$ matrix,

$$\mathbf{Z}_t = \begin{pmatrix} \mathbf{Z}_{t1} \\ \vdots \\ \mathbf{Z}_{tK} \end{pmatrix},$$

where entry Z_{tij} is the random number of individuals who move from state i to state j from time t to $t+1$.

Then, the general stochastic CM for $t \in \{1, \ldots, T\}$ is

$$\begin{aligned} \mathbf{X}_0^{CM} &= \mathbf{X}(0) \\ \mathbf{X}_t^{CM} &= \mathbf{Z}_{t-1}^T \cdot \mathbf{1}_K, \end{aligned} \tag{2.9}$$

where $\mathbf{1}_K$ is a vector of ones of length K. The number of individuals in state i at time t is equal to the number of individuals moving from state j to state i, summing over all j, from time $t-1$ to t. The law of mass action is what allows us to say $p_{ij}(t) \in [0, 1]$ for all $i = 1, \ldots, K$, and by design $\sum_{j=1}^{K} p_{ij}(t) = 1$. We are thus specifying a valid probability distribution. We let $\mathbf{X}^{CM} = \mathbf{X}(0)$, that is the initial values in both the deterministic and stochastic models are equal and known.

Stochastic AM. We can create a matching stochastic AM for the deterministic CM in Eq. (2.9). In keeping with our matrix notation, instead of letting an agent A_{tn} equal a scalar value, we now say

$$\mathbf{A}_{tn} = \mathbf{e}_k \in \{0, 1\}^K,$$

where the kth entry of the column vector $\mathbf{e}_i$ is one and the rest are zero. We say that an agent $\mathbf{A}_{tn} = \mathbf{e}_k$ is in state k at time t.

Given the correct initial values in each state, for an agent $\mathbf{A}_{tn}$, $n = 1, 2, \ldots, N$ an agent may move (or may not) from its current state i to state j based on a probabilistic draw $\mathbf{W}_{tni}$, a row vector of size K having

a Multinomial distribution,

$$\mathbf{W}_{tni} \overset{iid}{\sim} \text{Multinomial}\,(1, \mathbf{p}_i(t))) \quad \text{for } i = 1, \ldots, K.$$

Then $\mathbf{W}_{tn}$ is a matrix of size $K \times K$

$$\mathbf{W}_{tn} \overset{iid}{\sim} \begin{pmatrix} \mathbf{W}_{tn1} \\ \vdots \\ \mathbf{W}_{tnK} \end{pmatrix}.$$

In words, $\mathbf{W}_{tni}$ indicates which state agent n will move to from time t to $t+1$, given that agent n is currently in state i. The index n says that each agent within the same state moves according to an independent, identically distributed (iid) variable.

Agent n then updates according to

$$\mathbf{A}_{tn} = \mathbf{W}_{t-1,n}^T \mathbf{A}_{t-1,n}. \tag{2.10}$$

That is, agent n only moves from state i to j from time $t-1$ to t if and only if $\mathbf{A}_{t-1,n} = \mathbf{e}_i$ and $W_{tnij} = 1$. Equation (2.10) may make more intuitive sense, if we view it in terms of its transpose,

$$\mathbf{A}_{tn} = \left(\mathbf{A}_{t-1,n}^T \mathbf{W}_{t-1,n}\right)^T.$$

Then we see that because $\mathbf{A}_{t-1,n} = \mathbf{e}_i$, the only non-zero terms are obtained from the ith row of $\mathbf{W}_{t-1,n}$.

The aggregate total in the states, expressed as a column vector, is

$$\mathbf{X}_t^{AM} = \sum_{n=1}^{N} \mathbf{A}_{tn}. \tag{2.11}$$

We show the stochastic CM and stochastic AM have equivalent joint distributions in terms of the size of the states.

Theorem 2.4. *Fix an underlying deterministic CM with K states and known initial values and difference equations. Let the stochastic CM be described by Equation* (2.9) *and the stochastic AM be described by Equation* (2.11)*. Then*

$$\mathbf{X}^{CM} \overset{d}{=} \mathbf{X}^{AM}.$$

Proof. The main idea in this proof is due to the Multinomial/Multinomial relationship described in Equation (2.6) and a recurrence relation between the past step and the current step.

We begin with the stochastic AM specification in Equation (2.10) and (2.11) and show that this is equal in distribution to the stochastic CM specification in Equation (2.9),

$$
\begin{aligned}
\mathbf{X}_t^{AM} &\overset{d}{=} \sum_{n=1}^{N} \left(\mathbf{W}_{t-1,n}\right)^T \mathbf{A}_{t-1,n} \\
&\overset{d}{=} \sum_{n=1}^{N} \left(\mathbf{A}_{t-1,n}^T \mathbf{W}_{t-1,n}\right)^T \\
&\overset{d}{=} \sum_{n=1}^{N} \sum_{i=1}^{K} \left(\mathcal{I}\left\{\mathbf{A}_{t-1,n} = \mathbf{e}_i\right\} \cdot \mathbf{e}_i^T \mathbf{W}_{t-1,n}\right)^T \qquad (2.12) \\
&\overset{d}{=} \sum_{n=1}^{N} \sum_{i=1}^{K} \left(\mathbf{e}_i^T \left(\begin{array}{c} 0 \\ \vdots \\ \mathcal{I}\left\{\mathbf{A}_{t-1,n} = \mathbf{e}_i\right\} W_{t-1,ni} \\ \vdots \\ 0 \end{array} \right) \right)^T \qquad (2.13) \\
&\overset{d}{=} \sum_{i=1}^{K} \left(\mathcal{I}\left\{\mathbf{A}_{t-1,n} = \mathbf{e}_i\right\} \cdot \mathbf{e}_{t-1,i}^T \left(\begin{array}{c} 0 \\ \vdots \\ \sum_{n=1}^{N} \mathcal{I}\left\{\mathbf{A}_{t-1,n} = \mathbf{e}_i\right\} W_{t-1,ni} \\ \vdots \\ 0 \end{array} \right) \right)^T \qquad (2.14) \\
&\overset{d}{=} \sum_{i=1}^{K} \left(\mathbf{e}_i^T \left(\begin{array}{c} 0 \\ \vdots \\ Z_{t-1,i} \\ \vdots \\ 0 \end{array} \right) \right)^T \qquad (2.15) \\
&\overset{d}{=} \left(\mathbf{1}^T \cdot \mathbf{Z}_{t-1}\right)^T \\
&\overset{d}{=} \mathbf{Z}_{t-1}^T \cdot \mathbf{1} \\
&\overset{d}{=} \mathbf{X}_t^{CM}.
\end{aligned}
$$

Equation (2.12) to Equation (2.13) is due to the fact that the only draws that are relevant for an agent in state i, are the Multinomial draws in row i of $\mathbf{W}_{t-1,n}$. Since in Equation (2.14),

$$\sum_{n=1} \mathcal{I}\left\{\mathbf{A}_{t-1,n} = \mathbf{e}_i\right\} = X^{AM}_{t-1,i},$$

and since $W^T_{t,ni}$ are independent and identically distributed (iid), Equation (2.14) to (2.15) is due to the Multinomial/Multinomial equivalence described in Equation (2.6).

The proof is completed since the initial values are known at time $t = 0$ for both the CM and the AM, and we can use the recurrence relation for the proceeding steps.

□

Theorem 2.4 allows us to create stochastic CM-AM pairs given there is an underlying, "true" deterministic CM with transition matrix $\mathbf{D}(t)$. This means that on average, both our stochastic CM and AM will look like the true model, which can be useful if we want our models to have certain shapes.

2.4 Summary

In this chapter, we showed how AMs can fit into the CM statistical framework. We first showed an example of the Kermack and McKendrick deterministic SIR-CM with a corresponding stochastic CM-AM pair. We demonstrated through simulations that this CM-AM pair is jointly equal in distribution in terms of the number of individuals in the S, I, and R states. The main idea behind this equivalence is that the sum of N independent Bernoulli variables with a given probability p is equivalent in distribution to a Binomial draw with size n and probability p.

We then showed that given a deterministic transition matrix $\mathbf{D}(t)$, we can create equivalent stochastic CM-AM pairs. We proved this theorem with the use of the Multinomial/Multinomial equivalence, which is an extension of the Bernoulli/Binomial equivalence. The Multinomial/Multinomial equivalence is that the sum of N independent Multinomial variables drawn from a population of size 1 and probability $\mathbf{p}$ is equivalent in distribution to a Multinomial variable from a population of size N and probability $\mathbf{p}$.

The result of Theorem 2.4 is that we have an exact equivalence of CMs and AMs given an underlying $\mathbf{D}(t)$. There is no need for limits or asymptotics as suggested by Eubank et al. (2010).

This chapter focuses on stochastic models with underlying model shapes given by deterministic transition matrices $\mathbf{D}(t)$, which are directly analogous to the original K&M equations. Theorem 2.4 allows to put these deterministic, discrete time models into a stochastic framework, which allows us to account for noisy data. Theorem 2.4 relies on the independence of agents who are currently in the same state.

However, we do not need $\mathbf{D}(t)$ to create equivalent stochastic CM-AM pairs, nor do we need independence of agents. In Chapter 3, we explore conditions for equivalent CM-AM pairs when we relax or remove $\mathbf{D}(t)$ and allow for dependency of agents within the same state.

Chapter 3

General CM-AM equivalence

In this chapter, we relax the assumptions required in Theorem 2.4 required to create equivalent stochastic CM-AM pairs. Specifically, we examine the assumptions of independence of agents and the reliance on the underlying shape of the model given by a transition matrix $\mathbf{D}(t)$. Classically, epidemiological models were built with a set of deterministic difference (or differential) equations in mind such as those of Kermack and McKendrick (1927). Unsurprisingly, many stochastic models were developed to incorporate the deterministic difference equations, usually in terms of expected value of the stochastic model (Daley et al., 2001). In Chapter 2, we aimed to maintain the underlying shape of these deterministic difference equations by basing our stochastic CM-AM pairs on the underlying transition matrix $\mathbf{D}(t)$ to put it more in line with the K&M series of models.

In order to create equivalent CM-AM pairs in Chapter 2, we relied on having agents who occupy the same state as being independent from one another. However, in epidemiology we know that such an assumption is extremely suspect as there is evidence of sex, age, socio-economic factors in the spread of disease through a population (Koide and Seno, 1996). In this chapter, we relax the assumptions used in the previous model to create more general CM-AM pairs.

3.1 Base assumptions

Before relaxing assumptions for CM-AM pairs, we find it useful and natural to maintain the following assumptions:

1. An individual/agent may belong to only one state at a given time, which means the number of agents in states at a given time are non-negative integers. That is for $t = 0, \ldots, T$,

$$\mathbf{X}_t \in \mathbb{Z}_{\geq 0}^K.$$

2. A CM is characterized by

   (a) Homogeneity of individuals within states

   (b) Homogeneous interaction between states of individuals in susceptible and infectious states,

3. An agent update in an AM is necessarily a conditional Multinomial draw of size 1, namely

$$\mathbf{A}_{t,n}|\mathbf{A}_{t-1}, \mathbf{p}_{t-1,n} \sim \text{Multinomial}\left(1, \mathbf{p}_{t-1,n}\right).$$

   where $\mathbf{p}_{t-1,n}$ is possibly a random variable dependent on the states of other agents.

Assumption (1) may seem strange because it in fact excludes common deterministic models such as the K&M SIR-CM, which more often than not, gives fractional values for the number of individuals in a given state at a given time. However, since we are modeling discrete data, we find this assumption of integer values to be acceptable. Moreover, this does not mean that $E[\mathbf{X}_t]$ is a vector of non-negative integers.

Assumption (2) codifies the major simplification of CMs – homogeneity within states and homogeneous mixing. This is what allows us to essentially consider an individual within a state to be the same as any other individual within a state. Moreover, homogeneous mixing tells us that susceptible and infectious agents are equally likely to interact and hence become infected.

Assumption (3) gives the structure of the most basic AM. An agent, given its previous state and probability of moving, has some chance of moving to another state. This probability $\mathbf{p}_{t,n}$ may itself be a random variable. Since there is a bounded number of states to move to, then this new agent state, conditioned on the previous states and $\mathbf{p}_{t,n}$ is necessarily a Multinomial draw. As such, we can view conditional Multinomial draws as the foundation of AMs.

The assumptions that we are relaxing include (1) requiring individuals within states to be independent from one another, (2) requiring a CM to be given by a Multinomial draw, and (3) requiring both the CM and AM movements from one state to the next to be determined by a deterministic transition matrix $\mathbf{D}(t)$. Relaxing assumption (1) allows us to examine what it means for a group of individuals/agents to be "homogeneous," and we find that being independent from one another is a sufficient but not necessary condition. Relaxing assumption (2) allows us more freely to model how a number of individuals within a group move from one state to the next, which allows us to use models including Poisson draws, Binomial draws where the probability of moving is a Beta random variable and other parametric or non-parametric specifications (Daley et al., 2001; King et al., 2015). Relaxing assumption (3) allows modellers to rely on functions other than K&M-style deterministic difference equations to specify the shape of the curve of the number of individuals in each state over time.

3.2 CMs have an AM pair

Given a (stochastic) CM, we can create a corresponding (stochastic) AM that has equivalent distributions in terms of the number of individuals in each state.

Let there be a CM with initial states $\mathbf{X}_0^{CM} = \mathbf{x}_0$ with K states given by

$$\mathbf{Z}_{t,k} | X_{t-1,k}^{CM} \sim F_{t,k} \tag{3.1}$$

where $\mathbf{Z}_{t,k}$ is a vector of non-negative integers of length K such that $\mathbf{Z}_{t,k}\mathbf{1} = X_{t-1,k}^{CM}$. In words, random variable $\mathbf{Z}_{t,k}$ is a random draw from the number of individuals at time $t-1$ in state k, $X_{t-1,k}^{CM}$ and partitioning them into the K states at time t based on CDF $F_{t,k}$.

The equivalent AM is given by the following,

$$\begin{aligned} \mathbf{A}_{0,n} = \mathbf{e}_k \text{ if } \sum_{j=1}^{k} \mathcal{I}\left\{j > 1\right\} X_{0,j-1} < n \leq \sum_{j=1}^{k} X_{0,j} \text{ for } k = 1, \ldots, K \\ \mathbf{A}_{t,n} | \mathbf{A}_{t-1}, \mathbf{p}_{t-1,n} \sim \text{Multinomial}\left(1, \mathbf{p}_{t-1,n}\right), \end{aligned}$$

where $\mathbf{p}_{t-1,n}$ is constructed so that $X_{t,k}^{AM} \overset{d}{=} X_{t,k}^{CM}$. To do this construction, for each state k, we take realizations of $\mathbf{Z}_{t,k}$ and permute the indices of the agents currently in state k, and assign the first $Z_{t,k,1}$ permuted-order agents to state 1 at time t, the next $Z_{t,k,2}$ permuted-order agents to state 2, and so on until we have assigned all our agents for all previous states k. Mathematically, this has complicated notation. First let $\mathcal{J}_{t,k} = \{n : \mathbf{A}_{t,n} = \mathbf{e}_k\}$ be the set of indices of agents in state k at time t. Let $\sigma_{t,k}$ be a function

$$\begin{aligned} \sigma_{t,k} : \mathcal{J}_{t-1,k} &\to \{1, \ldots, X_{t-1,k}\} \\ n &\to \sigma_{t,k}(n) \end{aligned}$$

that maps index n to a permuted value between 1 and $X_{t-1,k}$. Finally, let $\sigma_{t,k}$ itself be a random permutation drawn from all such permutations of size $X_{t-1,k}$, which we call $G_{t,k}$. Then $\mathbf{p}_{t-1}$ results from first obtaining realizations of $\mathbf{Z}_{t,k}$ and $\sigma_{t,k}$,

$$\begin{aligned} \mathbf{Z}_{t,k} | X_{t,k}^{AM} &\sim F_{t,k} \\ \sigma_{t,k} &\sim G_{t,k} \\ p_{t-1,n,i} &= \begin{cases} 1 & \text{if } \sum_{j=1}^{i} \mathcal{I}\left\{j > 1\right\} Z_{t,k,j-1} \leq \sigma_{t,k}(n) \leq \sum_{j=1}^{i} Z_{t,k,j} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.2}$$

By construction $\mathbf{X}^{CM} \stackrel{d}{=} \mathbf{X}^{AM}$, since we are selecting the number of agents/individuals moving from one state to the next from the same random variable as in the CM, $\mathbf{Z}_{t,k}$. The permutation of the agents is required because otherwise we would be able to distinguish between them, violating the assumption of homogeneity within groups.

Example 3.2.1. (SIR-AM with and without permuting indices) In this example, we demonstrate how two AMs can have equivalent distributions in terms of the number of individuals in each state but have distinguishable agents.

Let there be a stochastic SIR-CM with model parameters shown in Figure 2.2: $N = 1000$, $S(0) = 950, I(0) = 50)$, $\beta = 0.10$, and $\gamma = 0.03$. We would like to make a corresponding SIR-AM. One way to generate an AM is to use the process described in Eq. (3.2), which we will refer to as the "permuted" version. This is one of many ways to generate a corresponding AM for a given SIR-CM. We will also explore one other way.

Instead of permuting the indices at each time step, we instead could assign the first $Z_{t,i,1}$ agents who are currently in state i to be in state 1, the next $Z_{t,i,2}$ agents who are currently in state i to state 2, and so on. An analogy to this is that we have a stack of agents in each state at time 0 where the agent at the bottom of the stack has the smallest index. We use the SIR-CM at the next time step to select the *number* of agents who will move from one state to the next. We select *which* agents move by moving them one at a time: from the bottom of the stack of its current state and placing it on top of the stack of its next state. We will refer to this assignment of agents as the "non-permuted" version.

Both the permuted and non-permuted match the SIR-CM with respect to the number of agents in each state at each time. This is because the *number* of agents selected each at each time is based on a random variable that is equivalent in distribution to that of the random variable used in the SIR-CM. However, there is distinction in which agents move from one state to the next in the SIR-AM. For a SIR-CM to truly have a matching SIR-AM, we require the agents to be homogeneous within states and have homogeneous interaction. One way to express this requirement is through the property of exchangeability. That is for any permutation σ and for any any agent k_n who is currently in state k

$$P\left(A_{t,k_1}, A_{t,k_2}, \ldots, A_{t,k_N} | \mathbf{A}_{t-1}\right) = P\left(A_{t,\sigma(k_1)}, A_{t,\sigma(k_2)}, \ldots, A_{t,\sigma(k_N)} | \mathbf{A}_{t-1}\right). \tag{3.3}$$

This distinction is important because it can affect parameter estimation, especially in terms of uncertainty estimation. We demonstrate this distinction with the examples shown in Figures 3.1 and 3.2.

To recap, we design two AMs that have equivalent distribution in terms of the number of individuals in each state. In the first AM, the agents are made indistinguishable from one another by permuting the order of the agents at each time step before moving agents from one state to another. In the second AM, the agents are distinguishable as they move from one state to the next based on the original order they

were put in the state. We show this scenario in order to show the importance of fulfilling the assumption of homogeneity within states and homogeneous mixing when creating a corresponding AM for a given CM. If this assumption is not fulfilled, we can have different estimates of disease parameters.

We demonstrate this difference in estimation of disease parameters in Figure 3.1. We plot the empirical distribution of average time to infection for each of the 950 initially susceptible agents, which were simulated for L=1000 runs. Time to infection is associated with the parameter β (along with the total number of infectious individuals at each time step). Both the non-permuted and permuted AMs have the same sample mean, approximately 44 days until infection. However, it is clear that the non-permuted version has a much larger variance for time to infection than the permuted version. Intuitively, this can be explained that in the non-permuted version, the first agent is going to be infected in the shortest amount of time and the last agent in the longest time. In the permuted version, the initial ordering should not matter.

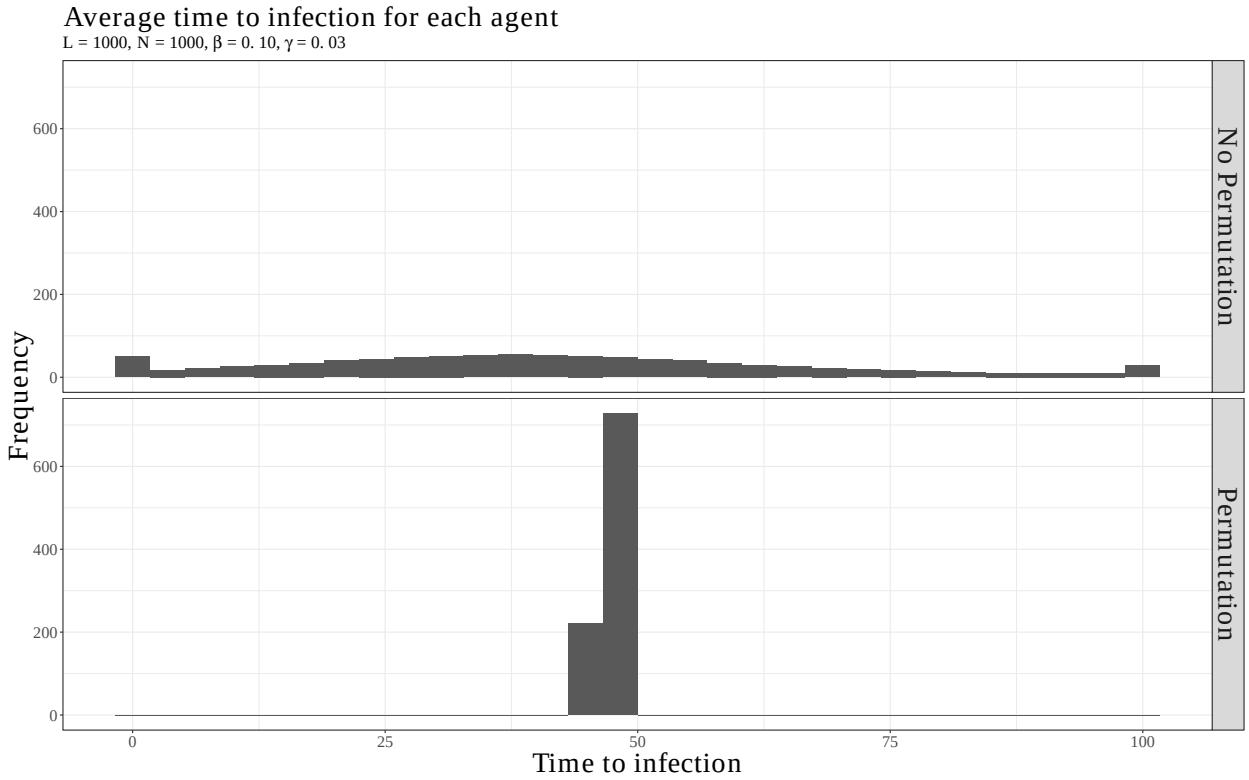

Figure 3.1: We plot the average time to infection over $L = 1000$ runs for each of 950 initially susceptible agents with $\beta = 0.10$ and $\gamma = 0.03$. In one set of simulations, we used the non-permuted version and the other used the permuted-version.

In fact, the importance of initial ordering can be seen in Figure 3.2. For the non-permuted AM, we see that Agent 1 (the agent at the bottom of the susceptible stack) has a much shorter time to infection than agent 201, who has a much shorter time to infection than Agent 401, and so on. The probability of infection from time 0 to time 1 decreases as n increases for each initially susceptible agent $n = 1, \ldots, 950$.

From this example, we see that the property of exchangeability shown in Eq. (3.3) is not being met for the non-permuted version. Conversely, for the permuted AM, there is no distinguishable difference among the distribution of time to infection for the different agents (see Figure 3.2 (bottom)), meaning that each initially susceptible agent has the same probability of being infected at each time step.

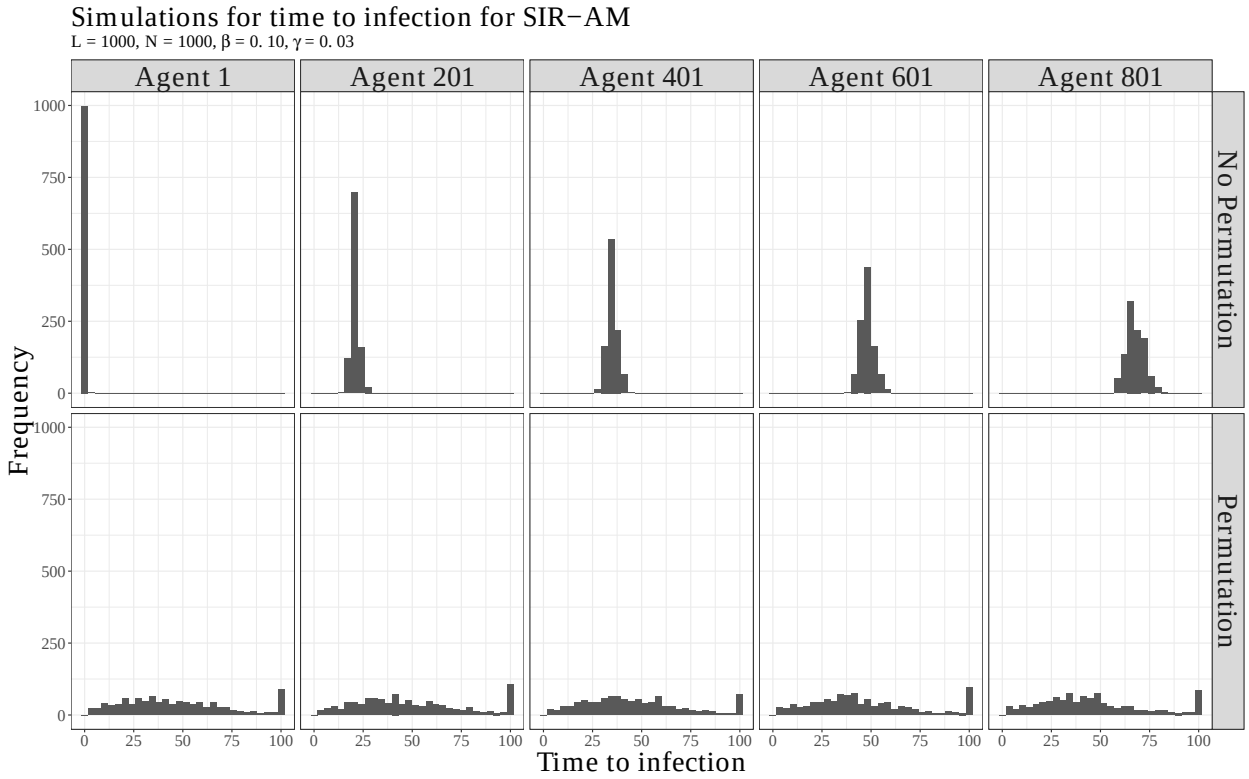

Figure 3.2: We plot the time to infection over $L = 1000$ runs for 5 specific agents with $\beta = 0.10$ and $\gamma = 0.03$. In one set of simulations, we used the non-permuted version and the other used the permuted-version.

The takeaway is that if the AM we design to correspond to a CM does not have homogeneity within states or homogeneous mixing, then we can easily misinterpret results from modeling. In the above example, if we used the non-permuted AM, we would conclude that time to infection has mean 44 days (95% CI: [34, 54]) days compared to the permuted AM where we would conclude time to infection has mean 44 days (95% CI: [40, 48]).

Example 3.2.2. (Lock-step) In this example, we relax the assumption that agents within the same states are independent from one another. Agents in this setting will instead move together in "lock-step." By lock-step, we mean that if one agent moves from one state to another from time $t-1$ to t, then the rest of the agents within the same state at time $t-1$ will also move to that new state. Instead of being independent of one another, agents now move through the epidemic in groups.

We again illustrate this concept with the SIR disease-level states, this time a subset of the system, which we call the S^2IR^2-system, shown in Figure 3.3. The S^2IR^2-system partitions the susceptibles into two separate states and the recovered individuals into two separate states. The susceptible individuals may move to a single infectious state and finally into one of two separate recovered (or removed) states.

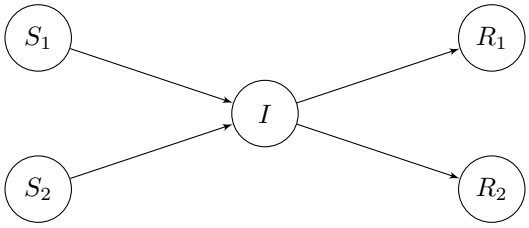

Figure 3.3: Graphical depiction of how individuals within a S^2IR^2-system move through states.

A deterministic S^2IR^2-CM is given by the following set of discrete-time, first order difference equations where the population size N is fixed,

$$\begin{cases} \frac{\Delta S_1(t)}{\Delta t} &= -\frac{\beta_1 S_1 I}{N} \\ \frac{\Delta S_2(t)}{\Delta t} &= -\frac{\beta_2 S_2 I}{N} \\ \frac{\Delta I(t)}{\Delta t} &= \frac{(\beta_1 S_1 + \beta_2 S_2)}{N} I - (\gamma_1 + \gamma_2)\, I \\ \frac{\Delta R_1(t)}{\Delta t} &= \gamma_1 I \\ \frac{\Delta R(t)}{\Delta t} &= \gamma_2 I \end{cases} . \tag{3.4}$$

In words, we have two distinct susceptible states (e.g. males and females) who have (possibly) different rates of infection, β_1 and β_2. However, once infectious, the two groups become indistinguishable from one another. Individuals within the infectious state then may either move into one of two "R" states (e.g. dead or recovered) at rates γ_1 and γ_2, respectively. To complete the model specification, we assume all initial states are known.

If we want to make equivalent stochastic CM-AM pairs where the agents in the AM are independent and mimic, on average, the shape of the curve of the difference equations, we can apply Theorem 2.4. Here, though, we would like to look at a "lock-step" situation, where individuals occupying the same state move together at the same time. In contrast to the previous section, the two susceptible groups of individuals may (or may not) be "locked" together in the I state. Once locked, the individuals within a state cannot unlock. That said, it is possible for the groups of individuals in the two initial susceptible states to remain unlocked throughout the entire epidemic. An alternative way to think of the lock-step model is weighting on the states that appears when determining the $\mathbf{p}_{tn}$ vectors but does not effect the number of agents transitioning.

For completeness, we provide a simulation of the stochastic S^2IR^2 CM-AM with the underlying set of discrete-time, first order difference equations described in Equation (3.4). Here, we set $\beta_1 = 0.25$, $\beta_2 = 0.50$, $\gamma_1 = 0.05$, $\gamma_2 = 0.10$. Additionally, $N = 1000$, $T = 50$, and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. Both the CM and AM were run for 5000 simulations.

In Figure 3.4, we plot the average percentage of individuals at each time step in each of the states for the CM (top) and the AM (bottom). Again, we see that the means for each percentage of individuals within a state at a given time look the same for both the CM and AM. For these parameters, we see, on average, nearly all individuals initially in $S_2(0)$ are infected but not so for the individuals initially in $S_1(0)$, which corresponds to $S_1(0)$ having a smaller infection rate than $S_2(0)$. Similarly, groups of individuals are much more likely to join the $R_2(t)$ recovered state than the $R_1(t)$ recovered state, about doubly so. However, we note that it is not so much the underlying deterministic CM that is important but rather, the fact that given the same underlying model that the stochastic CM and stochastic AM are equivalent in distribution. Figure 3.4 shows us that both the CM and AM have the same number of individuals in each state, on average.

In Figure 3.5, the corresponding sample variance of the number of individuals in each state at each time step are plotted for each of the models, again for 5000 total simulations. We see in the simulation, that the sample variances are similar although not exactly the same, which is similar to the case of the SIR-system, which also did not have exactly equal variances. This is evidence that the variances converge more slowly than the expectations at each time step. Additionally, the lock-step stochastic SIR is associated with larger variances of the states due to the binary nature of either all individuals within a group moving to the next state or staying in the current state. Given enough simulations, we would see these two lines completely overlap. This figure shows that the CM and AM also have the same variance in the number of individuals in each state.

Finally, we examine the distribution of sample paths in Figures 3.6-3.10 in order to see if the entire distributions of number of individuals in each state are the same for the CM and AM. In Figure 3.6, we have plotted the sample paths of the five states for both the stochastic CM (left) and stochastic AM (Right). In Figure 3.6, it is not clear that the distributions of the sample paths are the same, due to the step-wise movements of the lock-step model. As a result, the horizontal lines in the graphs are under-emphasized (as we cannot see many plotted on top of one another) and the near-vertical movements are overemphasized (as we see every transition even if occurred only once).

To better see that the distributions of the sample paths are the same for both the stochastic CM and stochastic AM, we present Figures 3.7-3.10. In Figure 3.7, for the two susceptible states, S_{1t} and S_{2t}, groups of individuals may transition at most once per simulation to the infectious state. The label in gray denotes how many individuals are moving at each time step. We plot the distribution of times of transition for the CM (left) and AM (right). We see that these distributions of times seem to be the same for the CM and AM for both susceptible states.

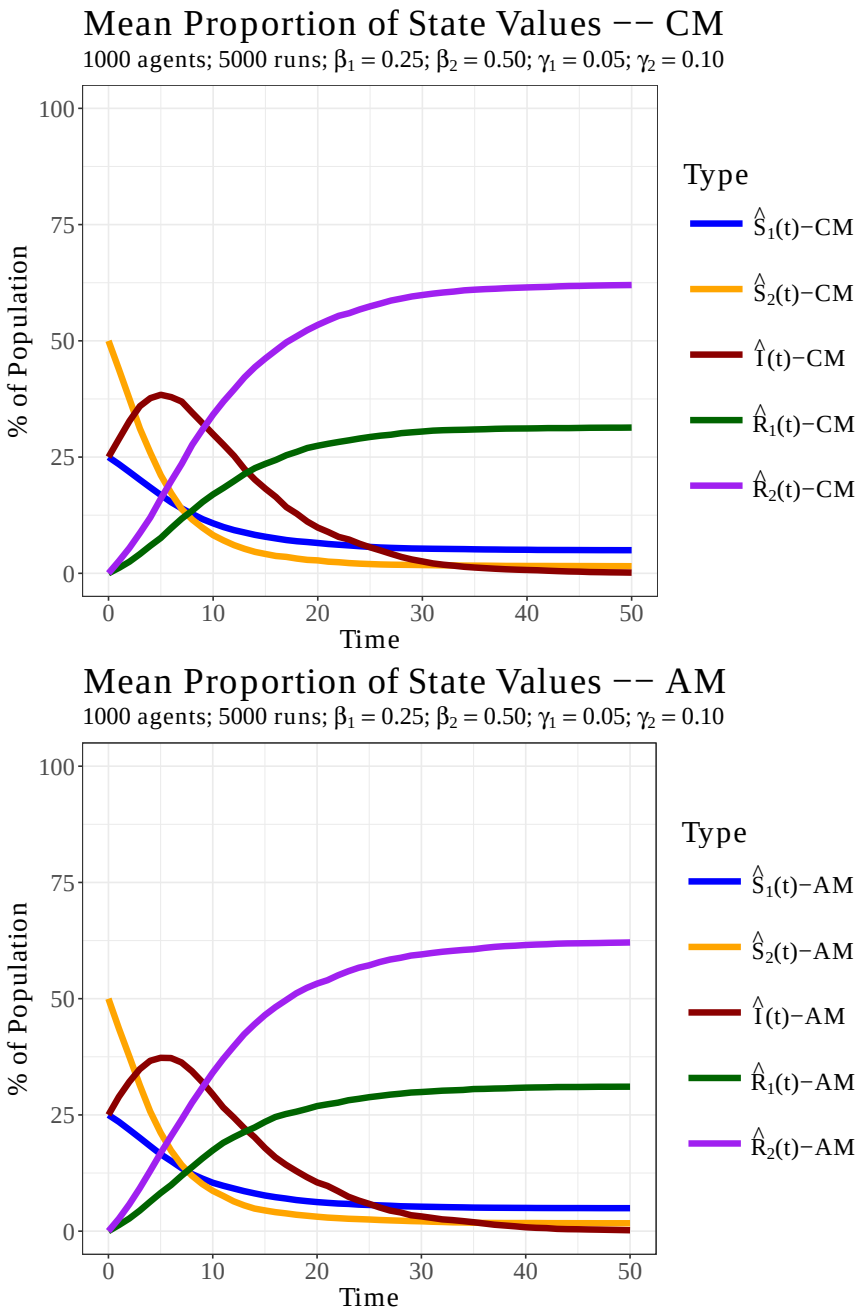

Figure 3.4: Top: CM approach. Bottom: AM approach. We plot the average percentage of individuals at each time step in each of the five states, $\hat{S}_1$, $\hat{S}_2$, $\hat{I}$, $\hat{R}_1$, and $\hat{R}_2$. We set $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\gamma_1 = .05$, and $\gamma_2 = 0.10$. Additionally, $N = 1000$ and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. Each model was run for a total of 5000 simulations.

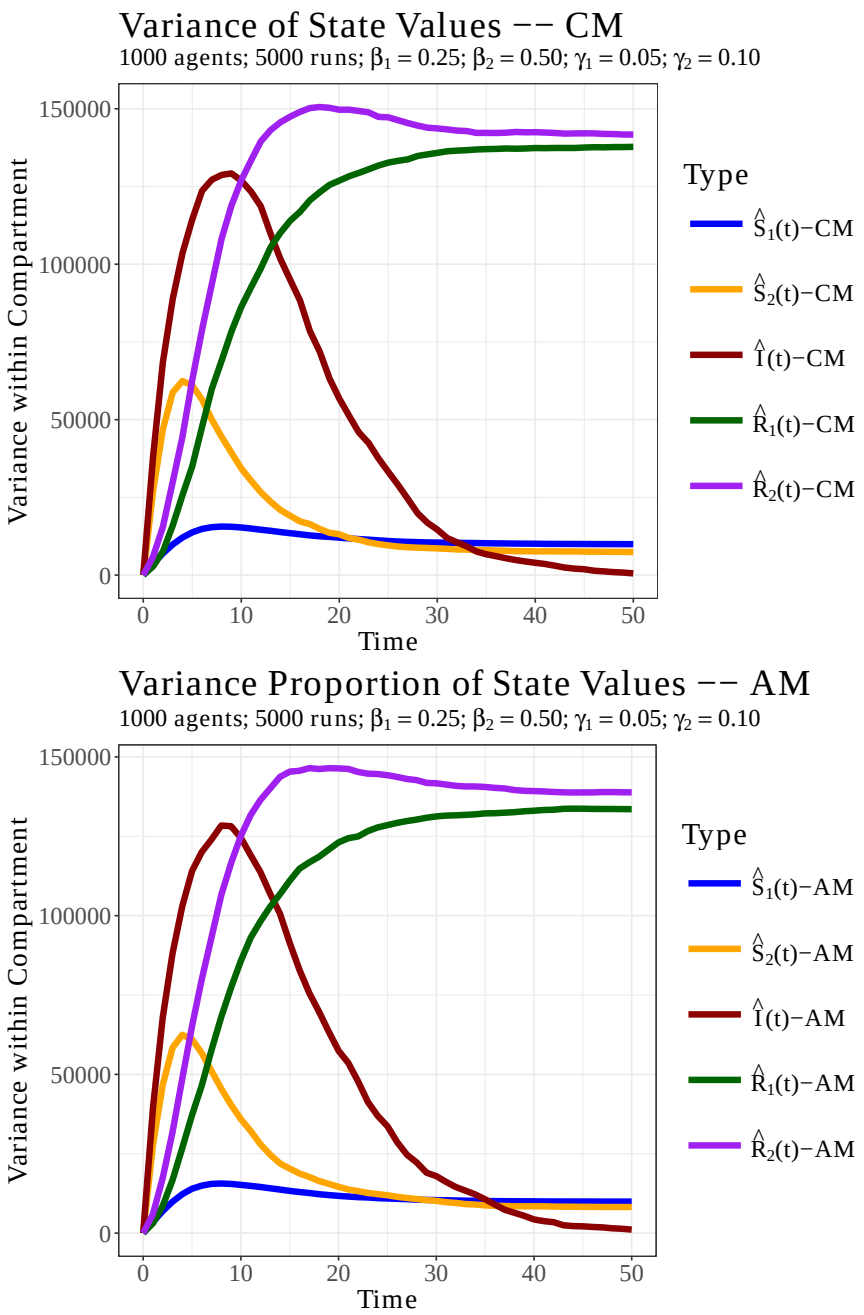

Figure 3.5: Top: CM approach. Bottom: AM approach. We plot the variance of individuals at each time step in each of the states. We set $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\gamma_1 = .05$, and $\gamma_2 = 0.10$. Additionally, $N = 1000$ and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. Each model was run for a total of 5000 simulations.

In Figure 3.8, we plot the distribution of transition times for the stochastic CM (left) and stochastic AM (right) for the infectious state I_t. There are seven total combinations of values of change of the number infectious, depending on whether groups are locked together or not. One of the most common cases of change (-250) is when the initial infectious individuals recover before infecting anyone else or when initially infectious individuals both infect the first susceptible state and recover in the same time step. Also commonly, we see changes of 250 and 500, which means that the initially infectious individuals are infecting one of the two susceptible states before recovering. We see instances of where one group of susceptibles is locked together with the initial infectious due to instances of changes of -500 and -750, respectively. In these simulations, we rarely see both groups of susceptibles locked together while the group of initially infectious individuals do not recover (change of 750), and thus, even more rarely see a change of -1000, when all three groups would move from the infectious state to one of the recovered states. We see that the distributions of change of individuals at each time steps seem to be the same for both the stochastic CM and stochastic AM.

In Figures 3.9 and 3.10, we plot the distribution of transition times for the stochastic CM (left) and stochastic AM (right) for the recovered states $\hat{R}_1$ and $\hat{R}_2$, respectively. Like in the infectious state, groups may be locked together to have different values of changes into the recovered states. We see that in many cases, the group of infectious individuals (of size 250) recover before infecting either group of individuals in susceptible states, which is why the percent of change of 500, 750, and 1000 is small compared to changes of 250. Again, we see that the distributions of the states for stochastic CM and stochastic AM look the same for both R_1 and R_2.

Overall, Figures 3.7-3.10 show that the distribution of transition times for groups of individuals from one state to another are equivalent in both the stochastic CM and stochastic AM. Since the distribution of times is derived from the number of individuals in each state at each time step, it is evidence that the CM and AM have equivalent in distributions in terms of the number of individuals in each state.

This example shows agents do not need to be independent from one another to have equivalent CM-AM pairs. We have demonstrated two extreme cases of dependency of individuals: (1) when the agents are completely independent from one another and (2) when the agents are completely dependent on one another (lock-step).

In reality, we expect the agents to have some dependency structure in between the two extremes. The independent case and lock-step cases are important in that the AM versions fulfill the requirement of the agents being homogeneous within states and having homogeneous mixing.

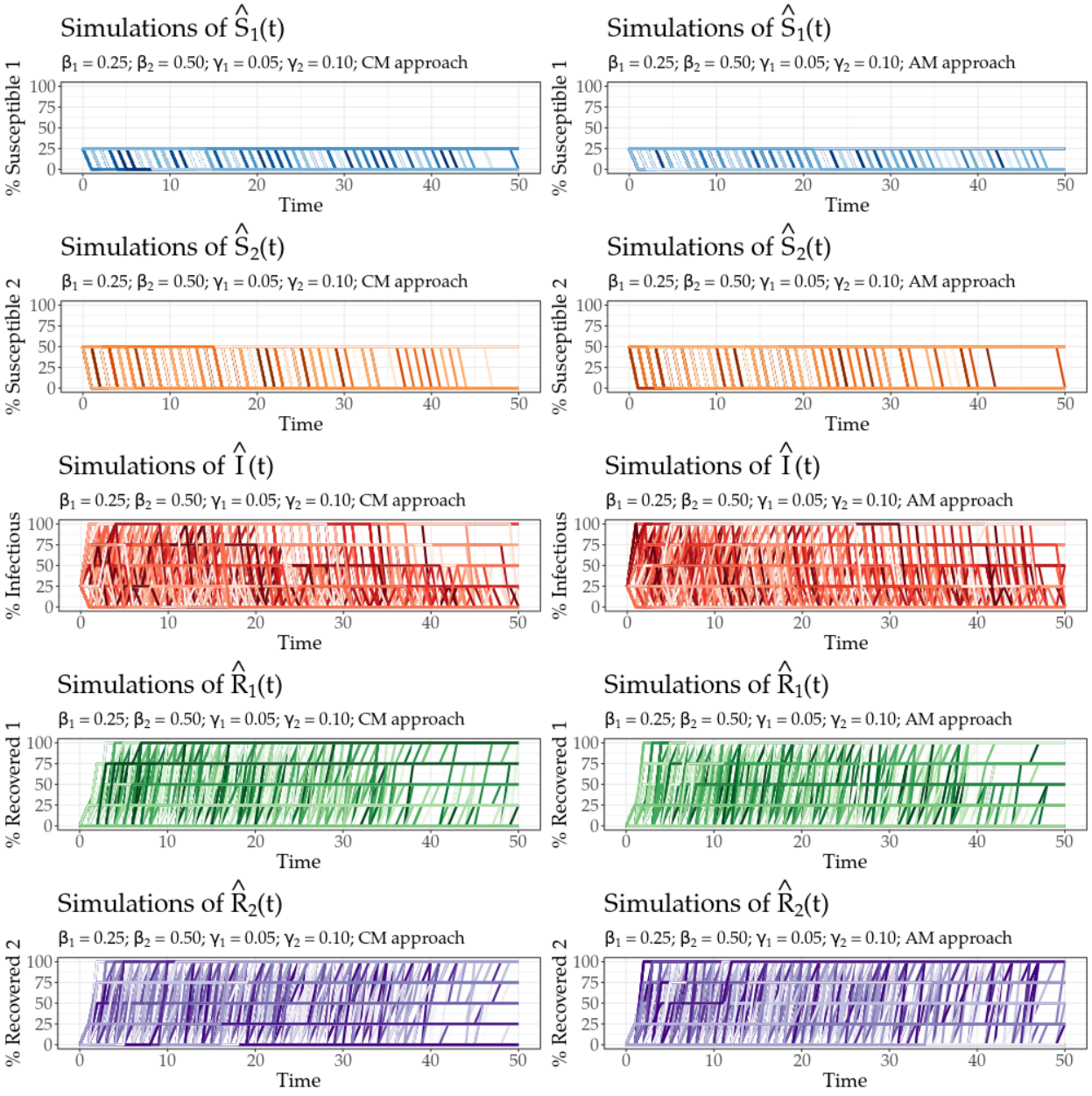

Figure 3.6: Top: CM simulation. Bottom: AM simulation. We plot the sample paths of the percent of individuals within the five states at each time step t. There are 5000 sample paths for each state. Here, we set $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\gamma_1 = .05$, and $\gamma_2 = 0.10$. Additionally, $N = 1000$ and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. Each model was run for a total of 5000 simulations.

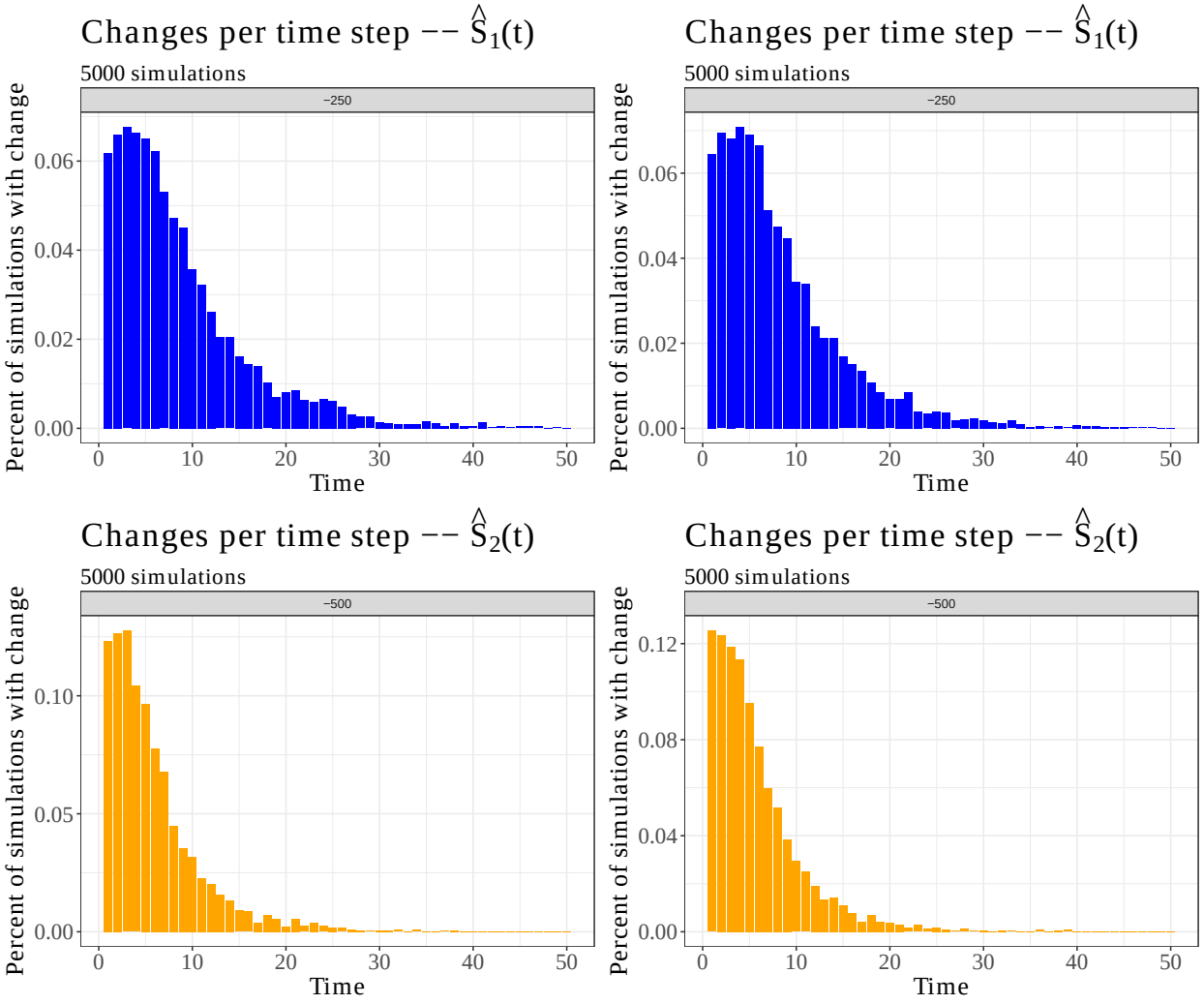

Figure 3.7: Left: CM simulation. Right: AM simulation. In the lock-step, stochastic S^2IR^2 CM and AM, the groups of susceptible individuals in states $\hat{S}_1(0)$ and $\hat{S}_2(0)$ have a chance to transition to the infectious state $\hat{I}(t)$ at each time step. This transition will happen at most once since groups of individuals are locked together. We plot the percent of transitions at time t for the 5000 simulations. The label in gray is the change of individuals, meaning the susceptible states may move all 250 or 500 individuals, respectively, at each time step to the infectious state. Here, we set $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\gamma_1 = .05$, and $\gamma_2 = 0.10$. Additionally, $N = 1000$ and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. We run both models 5000 times.

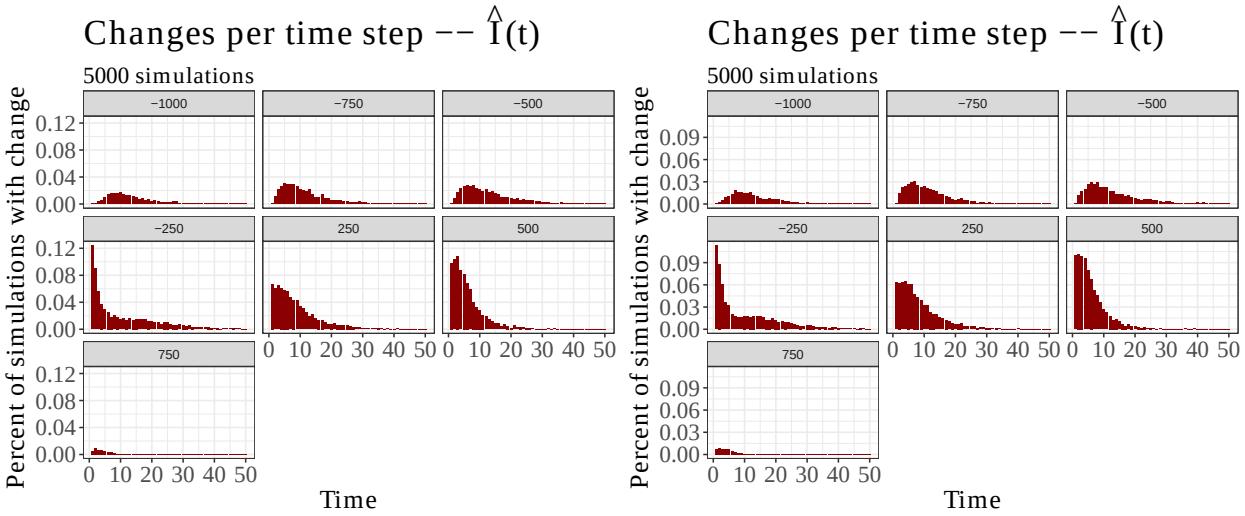

Figure 3.8: Left: CM simulation. Right: AM simulation. In the lock-step, stochastic S^2IR^2 CM and AM, the groups of individuals are moving both into and out of the infectious state $\hat{I}(t)$. We plot the percent of transitions at time t for the 5000 simulations. The label in gray is the value of the change of individuals within the $\hat{I}(t)$ state at each time step. For example, -250 indicates that the group of initially infectious individuals of state $\hat{I}(0)$ recover or, less commonly, the group in state $\hat{S}_1(0)$ has moved to the infectious state at the same time the group of initially infectious individuals recover. Here, we set $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\gamma_1 = .05$, and $\gamma_2 = 0.10$. Additionally, $N = 1000$ and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. We run both models 5000 times.

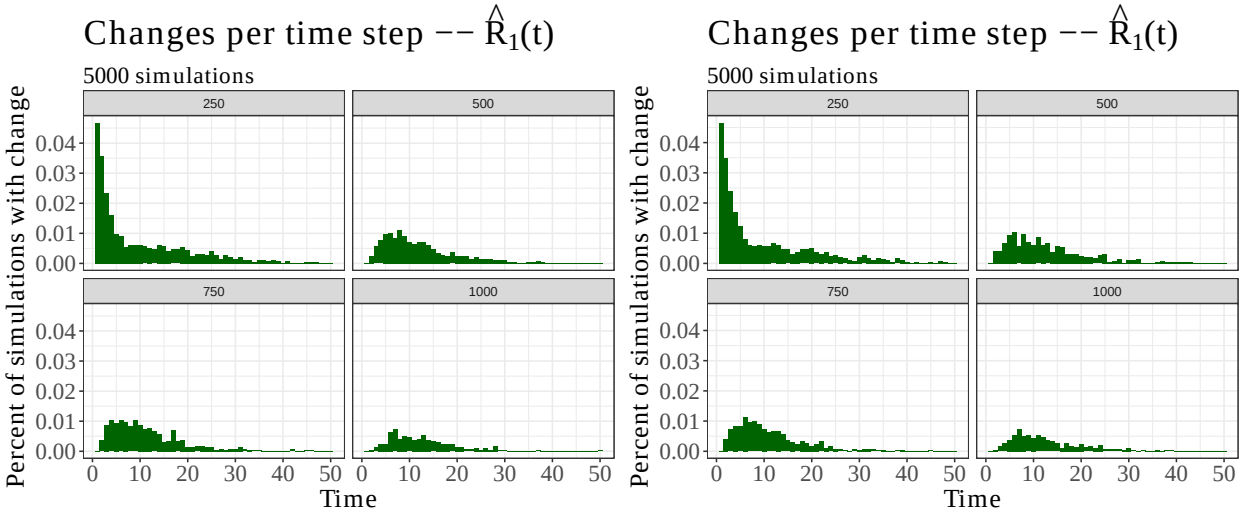

Figure 3.9: Left: CM simulation. Right: AM simulation. In the lock-step, stochastic S^2IR^2 CM and AM, infectious individuals in state $\hat{I}(t)$ have a chance to recover into one the recovered states, $\hat{R}_1$ and $\hat{R}_2$. We plot the percent of transitions at time t for the 5000 simulations. The label in gray is the value of the change of individuals within the $\hat{R}_1(t)$ state at each time step. For example, 250 indicates that the group of initially infectious individuals of state $\hat{I}(0)$ recover or the group of initially susceptible individuals of state $\hat{S}_1(0)$ recover at a different time than the the group of initially infectious individuals. Here, we set $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\gamma_1 = .05$, and $\gamma_2 = 0.10$. Additionally, $N = 1000$ and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. We run both models 5000 times.

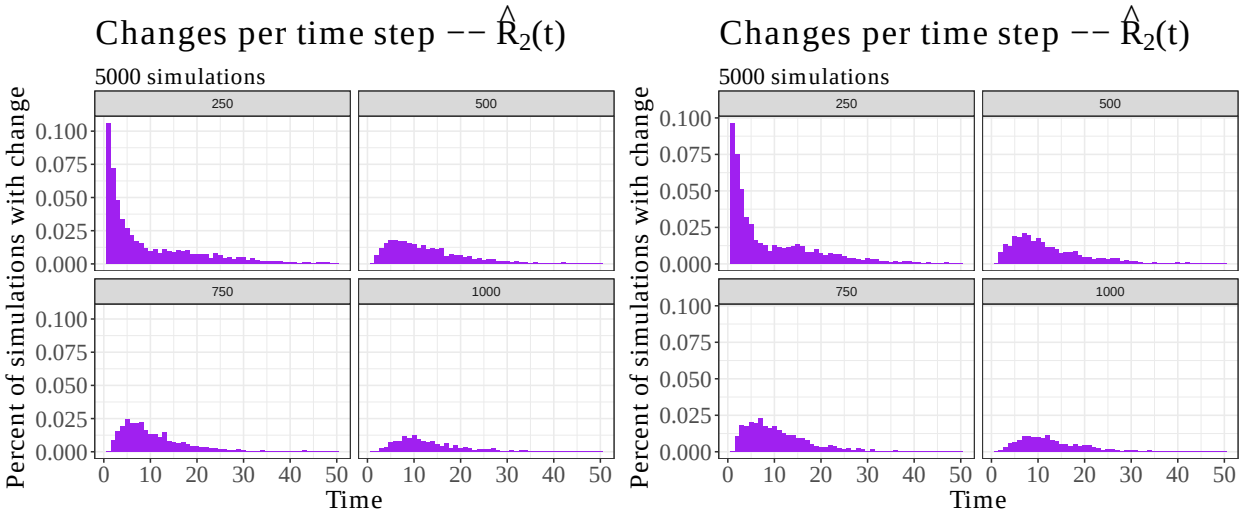

Figure 3.10: Left: CM simulation. Right: AM simulation. In the lock-step, stochastic S^2IR^2 CM and AM, infectious individuals in state $\hat{I}(t)$ have a chance to recover into one the recovered states, $\hat{R}_1$ and $\hat{R}_2$. We plot the percent of transitions at time t for the 5000 simulations. The label in gray is the value of the change of individuals within the $\hat{R}_2(t)$ state at each time step. For example, 250 indicates that the group of initially infectious individuals of state $\hat{I}(0)$ recover or the group of initially susceptible individuals of state $\hat{S}_1(0)$ recover at a different time than the the group of initially infectious individuals. Here, we set $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\gamma_1 = .05$, and $\gamma_2 = 0.10$. Additionally, $N = 1000$ and $(S_1(0), S_2(0), I(0), R_1(0), R_2(0)) = (250, 500, 250, 0, 0)$. We run both models 5000 times.

3.3 Any AM has an equivalent CM

While in Section 3.2 we showed that any CM has an equivalent AM, in this section we show that we can create a (stochastic) CM for a given (stochastic) AM. Let the AM have N agents and K total states. Let the agent update be given by a conditional Multinomial draw of size 1,

$$A_{t,n}|A_{t-1}, \mathbf{p}_{t,n} \sim \text{Multinomial}\,(1, \mathbf{p}_{t,n})$$
$$\mathbf{p}_{t,n}|A_{t-1} \sim F_{t,n}$$

The issue in having a CM mimic an AM is the problem of homogeneity within states and homogeneous interaction, which is closely related to the issue of dependent agents. In Section 3.2, we showed the lock-step example where agents within states are completely dependent upon one another. Because the agents are inseparable, it follows that they are homogeneous and have homogeneous mixing. Given a fixed number of states, it is difficult to fulfill the requirements of homogeneity and interaction mixing unless the agents are completely dependent or independent from one another. We can circumvent this issue instead by adding more states to the model.

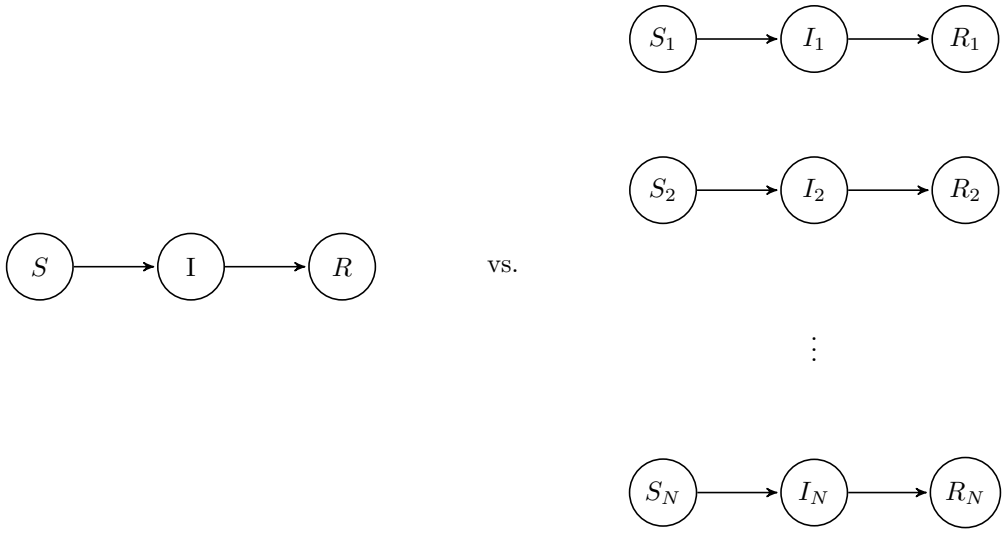

Figure 3.11: Two extreme CM depictions for a population of size N with three fixed disease-level states, SIR. On the left, there is one state for each disease-level state for a total of $K^* = 3$ states. On the right, there is one disease-level state for each agent for a total of $K^* = 3N$ states. For both models, we assume homogeneous mixing and homogeneity of individuals within states.

We partition the K states of the AM into $K^* = KN$ states. That is, our new states are

$$\mathbf{X}_t^{AM} = (X_{t,1}^{AM}, \dots X_{t,K}^{AM}, X_{t,K+1}^{AM}, \dots, X_{t,2K}^{AM}, \dots, X_{t,(N-1)K+1}^{AM}, \dots, X_{t,NK}^{AM}),$$

where the first K states are the states for the first agent and the next K states are for the second agent and so on. This partitioning is illustrated in Figure 3.11.

After partitioning the agents, simply let $\mathbf{X}_t^{CM} = \mathbf{X}_t^{AM}$. The CM and AM are jointly equivalent in distribution by construction. We need only show that this CM has homogeneity within states and homogeneous interaction. However, this is trivially true because there is at max one agent in every state at a given time.

The reader may wonder what is the point of this equivalence as it may seem more like a "trick" than a statistical result. This equivalence shows that CMs and AMs are exactly the same if we adjust the number of total states, again without having to look at asymptotic behavior. This equivalence allows us to identify 1) the minimal number of parameters to be estimated in either framework and 2) that the main feature that allowed us traditionally to discriminate between CMs and AMs (although they really are the same) is the number of total states.

3.4 Number of total states

Section 3.2 shows that every CM has an equivalent AM and Section 3.3 shows that every AM has an equivalent CM. The key differences between these sections is the total number of states we allow in the original model.

To better describe this issue, we define the concept of disease-level states. We define a disease-level state as any symptom or condition related to disease alone that is not associated with any demographic characteristics of agents, latent or otherwise. For example, the K&M deterministic SIR-CM has 3 disease-level states, S, I, and R. If, for example, males and females become infected at different rates β_1 and β_2, respectively, then it makes sense to partition the S disease-level state into two sub-states: S_M and S_F for a total of $K = 4$ states. This example is illustrated in Figure 3.12.

Figure 3.12: Depiction of two different models within the disease-level states SIR. For the left model, there are $K = 3$ total states. For the right model, there are $K = 4$ total states.

We refer to the model in Figure 3.12 (right) as the S^{2}IR-CM. The equations for the S^{2}IR-CM model are given by deterministic transition matrix $\mathbf{D}(t)$,

$$\mathbf{D}(t) = \begin{pmatrix} S_1(t-1) - \beta_1 \frac{I(t-1)}{N} & 0 & \beta_1 \frac{I(t-1)}{N} & 0 \\ 0 & S_2(t-1) - \beta_2 \frac{It-1)}{N} & \beta_2 \frac{It-1)}{N} & 0 \\ 0 & 0 & I(t) - \gamma I(t-1) & \gamma \\ 0 & 0 & 0 & R(t-1) \end{pmatrix} \tag{3.5}$$

If $\beta_1 = \beta_2 = \beta$ then the two susceptible states have the same rate of infection as in the original SIR model, and we can in fact model the population with $K = 3$ states.

In general, the question we need to focus on is how many total states K^* do we need to model a population given there are M-disease level states. We can bound K^* using the results of Sections 3.2 and 3.3. Given a fixed population N, then the minimum number of states required to model an outbreak of a disease is $M \leq K^* \leq MN$. In the next chapter, we will focus on techniques to determine whether K^* is a proper estimate of the total number of states and what that means in terms of our CM-AM pairs.

3.5 Summary

In this chapter, we showed that any CM has an equivalent AM and that any AM has an equivalent CM, regardless of whether there exists an underlying, deterministic transition matrix $\mathbf{D}(t)$.

In Section 3.2, we show there exists an AM pair for any CM with K states. We then examine the assumption of (in)dependence of agents. In the first example, we show that we can design an AM that is equivalent in distribution in terms of the number of individuals in each state to the CM and have the agents be distinguishable from one another. We show that having the agents being distinguishable from one another can, however, effect parameter estimates. In the lock-step example, we show an extreme case of dependency where agents within the same state are dependent on one another but still indistinguishable from one another.

In Section 3.3, we show how to create a matching CM for a given AM. In order to create a matching CM, we increase the number of total states in order to maintain homogeneity within states and homogeneous mixing of individuals.

Finally, in Section 3.4 we discuss the importance of the CM-AM equivalency, and in particular the number of total states used to model an outbreak. As the total number of states is directly related to the number of parameters needed to be estimated in either framework, it is key to find the minimum number of states that will adequately model an outbreak.

The equivalence between CMs and AMs in this section allow us to directly relate the parameters within the two frameworks. Thus, we can use established statistical techniques to estimate parameters for CMs and then transfer them over to the AM in framework. Once in the AM-framework, we can then examine hypothetical scenarios knowing how well it is docked to reality.

In the next chapter, we discuss techniques to determine K^* and to test whether our model, which may have a complex agent interaction structure, is a good fit to the data we observe.

Chapter 4

Model selection for CMs and AMs

In Chapter 3, we showed that if we adjust the number of total compartments K, then it is always possible to create an equivalent CM (AM) in terms of joint distribution of the number of individuals in the K states for a given AM (CM). The size of K is directly related to the number of parameters we need to estimate. For a fixed population N, and assuming every state is associated with at most one parameter θ_k, then there will be $K-1$ parameters to estimate. In order to estimate the simplest model possible (i.e. the one with the fewest number of parameters), we would like to determine K^*, the minimum number of total states required to adequately model an outbreak of a disease.

Once K^* is estimated, we can determine the "best" model with K^* states to select our CM-AM pair. We call the selected model a CM-AM pair to emphasize that the CM and AM are equivalent in terms of joint distribution, although we may view and analyze the two classes separately. Through the perspective of a CM, we can estimate our parameters and provide uncertainty estimates for them. We can then shift our perspective to that of an AM, where we can more easily explore hypothetical questions such as effects of shutting down schools or implementing isolation and quarantine strategies for our agents.

In this chapter, we explore methods to find K^*, the minimal number of total states needed to adequately model an outbreak. To this extent, we provide two novel diagnostic plots that are specific to the SIR disease-level states to aid in model selection. Additionally, we introduce a statistical investigation similar to that of Colizza et al. (2006) to determine K^* while working with the SI disease-level states. All coding, calculations, estimations, and simulations mentioned here are implemented in our `R` package `catalyst` available at https://github.com/shannong19/catalyst.

This chapter proceeds as follows. In Section 4.1 we present two novel diagnostics in the form of plots to determine whether our model is a good fit. Following that, in Section 4.2, we show how we can use statistical simulations to determine if we can simplify our model to one with fewer total states, K^*. Finally, in Section 4.3 we summarize the chapter.

4.1 SIR specific selection

SIR disease-level states were originally included in Kermack and McKendrick (1927) model, and models with SIR disease-level states and are still commonly used to assess and examine infectious disease outbreaks (Rizkalla et al., 2007; Zhao et al., 2013; Smith and Broniatowski, 2016; Mpeshe et al., 2017). The SIR model with $K^* = 3$ total states contains two disease parameters $\theta = (\beta, \gamma)$, the infection rate, and recovery rate, respectively. In this section, we present two novel diagnostic plots to be used to help assess the fit of the model and inform model selection. We show how a recent advancement in the deterministic SIR-CM can be used to formulate the deterministic K&M SIR-CM in a form more recognizable to statisticians which results in a plot useful in model assessment. Additionally, we present a separate ternary plot as a diagnostics tool for fitting SIR disease-state models.

4.1.1 The SIR and its relationship with linear regression

Harko et al. (2014) reduce the deterministic K&M 2-dimensional differential equations to 1-dimension. Specifically, they show that

$$S(t) = S(0) \exp\left\{-\mathcal{R}_0 \frac{R(t)}{N}\right\}, \tag{4.1}$$

where $\mathcal{R}_0 = \frac{\beta}{\gamma}$ is the reproduction number for the deterministic K&M SIR. Recall, $\mathcal{R}_0$ is the average number of secondary infections caused by a primary infection when the primary infection is introduced to a fully susceptible population (Anderson and May, 1992). In Eq. (4.1), the number of new susceptibles is based on exponential decay of the initial susceptible population where the decay is determined by the reproduction number $\mathcal{R}_0$ and the current percent of recovered individuals, $\frac{R(t)}{N}$. We adapt this deterministic equation to account for noise by replacing the deterministic variables with random variables (S_t, and R_t). Rearranging the terms and taking a log transformation, we have

$$\log\left(\frac{S_t}{S_0}\right) = -\mathcal{R}_0 \frac{R_t}{N}. \tag{4.2}$$

The equation in Eq. (4.2) is useful because it is in the form of standard linear regression, i.e. data that can be fit to a straight line through the origin.

We simulate SIR data from our Binomial movement model presented Eq. (2.2) and plot the results in Figure 4.1. We display the results of simulating 100 runs from the SIR-CM. In this figure, it is apparent that the variance of $\frac{R_t}{N}$ is not homoskedastic but increases as R_t increases. The best-fit weighted linear regression line is shown in red with a 95% prediction interval shown in black. To estimate the weights, we first estimate the $\hat{\sigma}_t^2 = V\left[\log\left(\frac{S_t}{S_0}\right)\right]$ as a straight line through origin as a function of R_t/N.

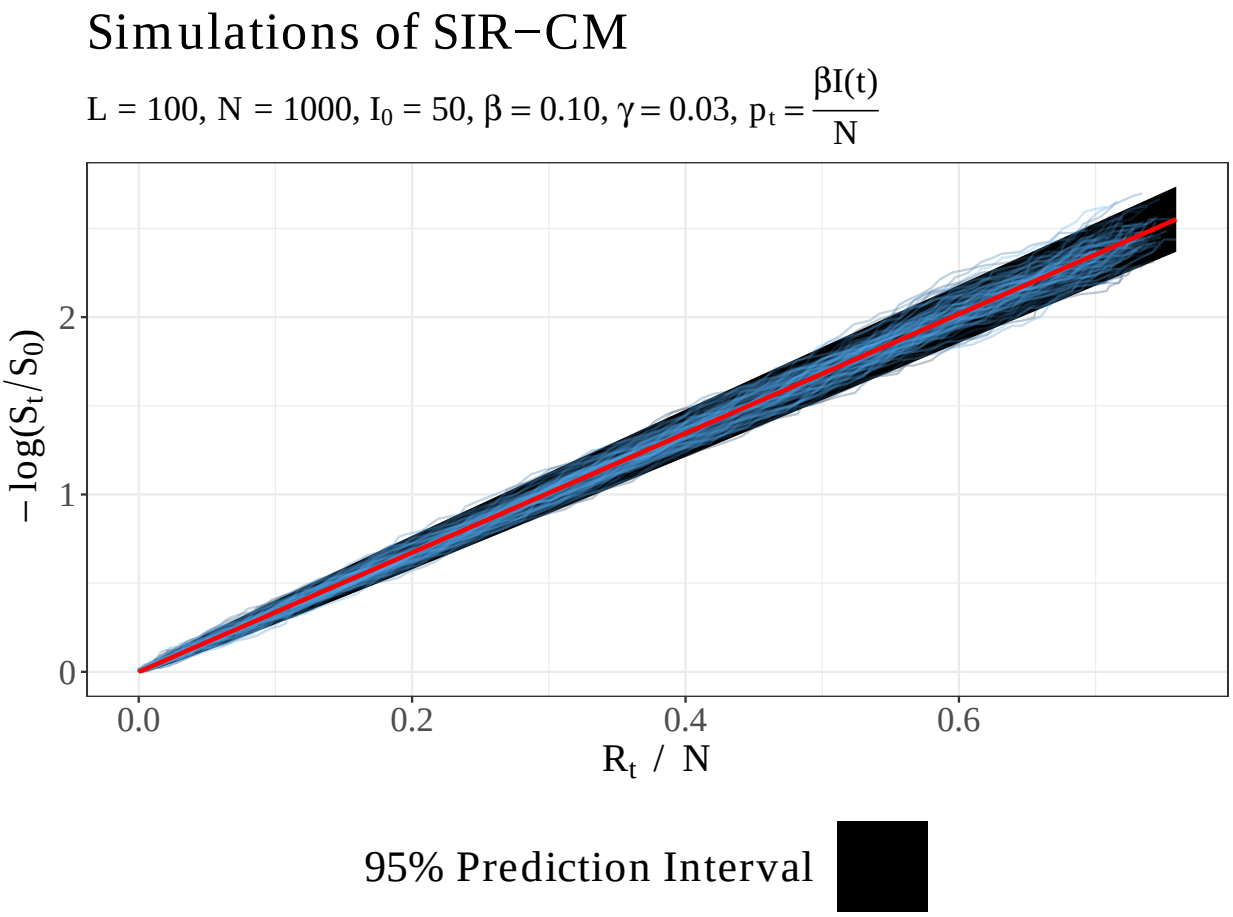

Figure 4.1: Simulations of SIR-CM with best-fit line (red) from weighted linear regression and 95% prediction interval from weighted least-squares linear regression.

For the weighted least squares regression model where we use the data points from all the different simulations to form our estimates, we estimate $\hat{\mathcal{R}}_0 = 3.400$ (95% CI: [3.397, 3.403]) whereas the true $\mathcal{R}_0 = 3.33$. Our estimate of the reproduction number along with its 95% CI does not include the true value of $\mathcal{R}_0$, but we see that the prediction interval covers the simulated data well.

Estimating the variance $\hat{\sigma}_t^2$ directly from the data is only a viable strategy when we have multiple sample paths, which typically does not occur in epidemic data. At best, we will observe (S_t, R_t) for $t = 0, \ldots, T$, meaning we will only have one sample path. We do, however, have an expression for the variance of S_t, in Eq. (2.4) and so can use the delta method to obtain an estimate for $V[\log(S_t/S_0)]$. Namely, when $S_t > 0$,

$$V[\log(S_t/S_0)] \approx \left[\frac{1}{S_t}\right]^2 V[S_t] \tag{4.3}$$

Given an estimate of $\hat{\beta}$, the average infection rate, which is necessary for the plug-in estimate of $V[S_t]$, we can then use weighted least squares regression with the weights as the inverse of Eq. (4.3). We randomly select one sample path from Figure 4.1 and fit a weighted regression line with the weights as the plug-in estimate for $1/V[\log(S_t/S_0)]$. This line and a 95% percent prediction interval are plotted in Figure 4.2. We estimate $\hat{\mathcal{R}}_0 = 3.40$ (95% CI: [3.31, 3.39]). We repeat the process for the other $L - 1$ sample paths and find that the average estimate of $\mathcal{R}_0$ is $\bar{\mathcal{R}} = 3.35$, and the true $\mathcal{R}_0$ is only covered 28% of the time. However, the coverage of the data (i.e. the sample paths are contained within the 95% prediction interval) is 95%.

Fortunately, the model is not very sensitive to our estimate of $\hat{\beta}$, which is required to estimate the plug-in variance $V[S_t]$ shown in Eq. (4.3). The data coverage is plotted as a function of $\hat{\beta}$ in Figure 4.3. We vary $\hat{\beta}$ between 0 and 1 and find the mean coverage of the data over $L = 100$ simulations, averaging over each of the the sample paths. The horizontal line is the 95% line, as we expect our 95% prediction intervals to cover the data 95% of the time. Our coverage is as expected unless we underestimate $\hat{\beta}$. Even when we underestimate $\hat{\beta}$, the coverage is over 94% except for values of $\hat{\beta} < \gamma$, the average recovery rate. The threshold $\hat{\beta} = \gamma$ is important because it is the value between observing an outbreak or not in the deterministic K&M SIR equations.

Overall, we see that plots like the one in Figure 4.2 can be used to assess whether a stochastic SIR-CM is a good fit to the observed SIR data, as we expect the slope of the best fit line to fit the points fairly well, although perhaps an overestimate of $\mathcal{R}_0$, and we expect approximately 95% of the observations to be covered in the prediction interval. In the future, we will investigate adding an additional intercept parameter to our model which would allow us to estimate $S(0)$ instead of assuming it is known.

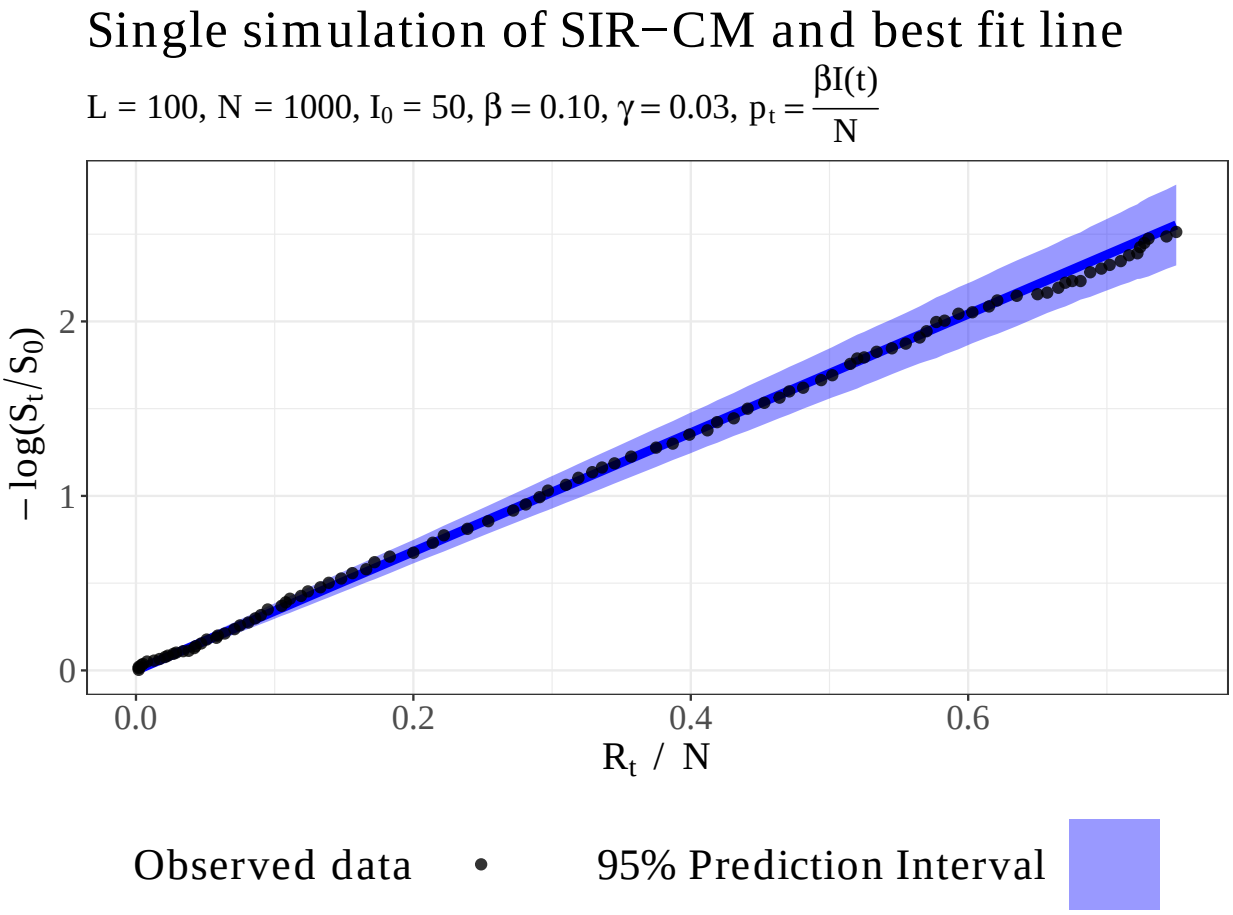

Figure 4.2: Single simulation of SIR-CM with best fit line (blue) and 95% prediction interval from weighted least-squares linear regression with the weights as the inverse of the plug-in estimate of Eq. (4.3). The slope of the line is also the estimate of $\mathcal{R}_0 = 3.40$

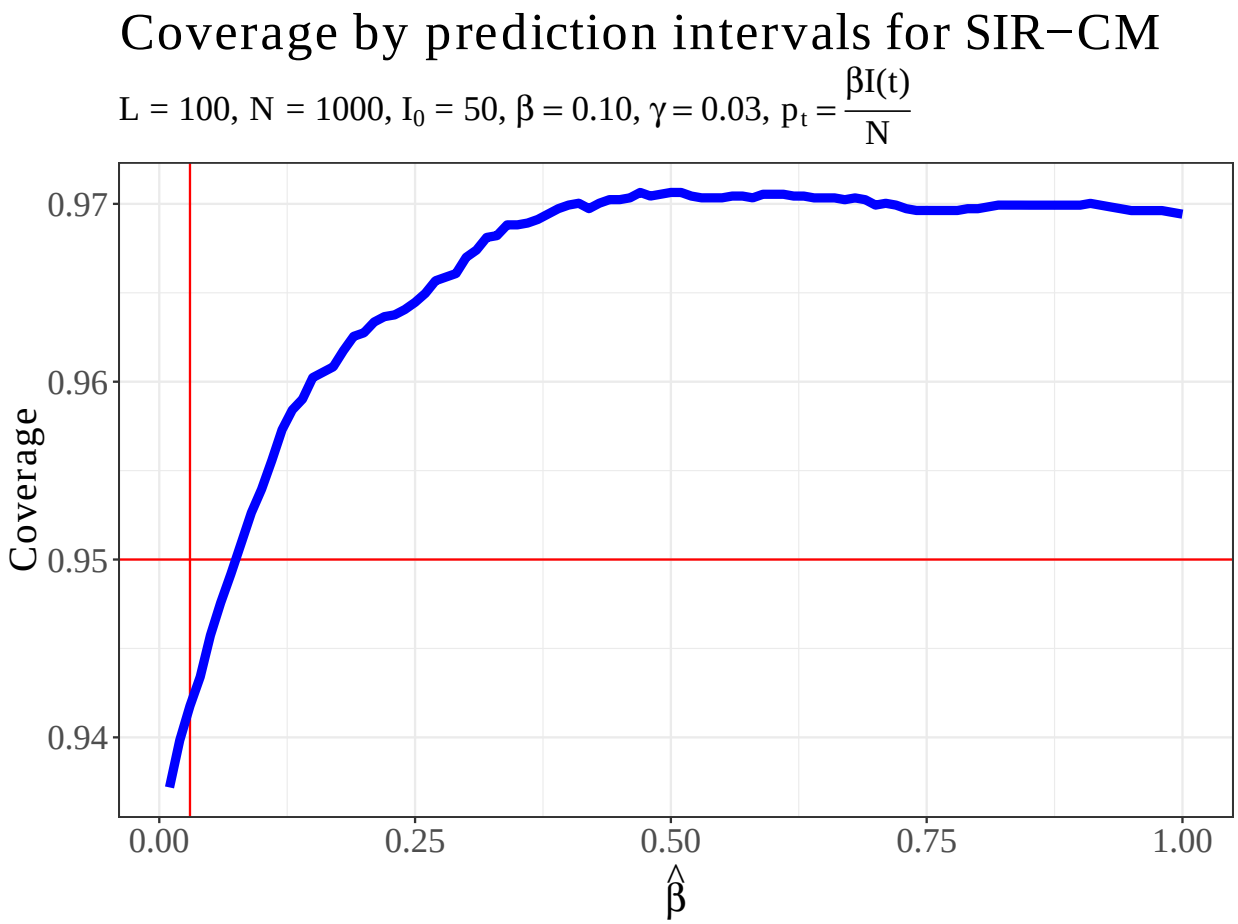

Figure 4.3: Coverage of data for our $L = 100$ SIR-CM simulations for different values of $\hat{\beta}$, which is used to estimate the weights for weighted linear regression.

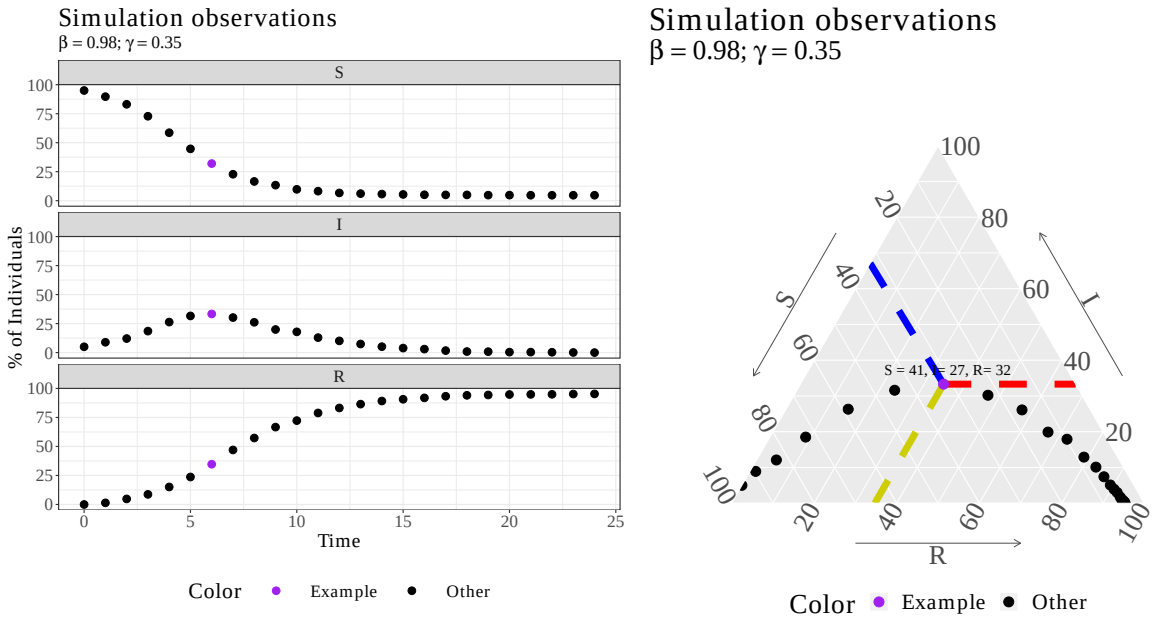

Figure 4.4: Observed SIR data simulated from the model in Eq. (2.2) with $\beta = 0.98$ and $\gamma = 0.35$. Left: % of individuals in state vs. time. Right: ternary plot of % in S, I, and R states. The point in purple is highlighted to show how the same point is represented in both plots.

4.1.2 Ternary plots

Since $S_t + I_t + R_t \equiv N$ and $S_t, I_t, R_t \in [0, N]$, plotting S_t, I_t, and R_t in three dimensions results in the observed data points lying in the plane constrained to the triangular region defined by $S_t, I_t, R_t \in [0, N]$. A ternary plot is then a natural way to display all three states simultaneously in two dimensions. Safan et al. (2006) present ternary plots for the SIS (which can be equivalently be displayed as an SIR model) as a way to show theoretical endemic equilibria from deterministic SIS differential equations. To our knowledge, the ternary plot has not been used as a visual diagnostic for model fitting and selection of an SIR model to data.

We demonstrate the concept here. We let $S(0) = 950$, $I(0) = 50$, and $R(0) = 0$. We simulate a set of observed data from the Binomial movement model in Eq. (2.2) with $\beta = 0.84$ and $\gamma = 0.30$ for times $t = 0, \ldots, T$. The observations are plotted in Figure 4.4 in the traditional view of % in state vs. time (left) and the ternary plot (right). We highlight a point in purple to show how the same point, in this case the point $(S_t = 41, I_t = 27, R_t = 32)$ is displayed in each plot. We see that the visualization the outbreak in the ternary plot is condensed to one point corresponding to (S_t, I_t, R_t). Although we do lose the time dimension in the ternary plot, we can visualize time in the ternary plot by using a color gradient for the observations. The ternary plot maintains primary features of the visualization such as monotonicity of the S and R states along with visibility of the peak of the infectious curve.

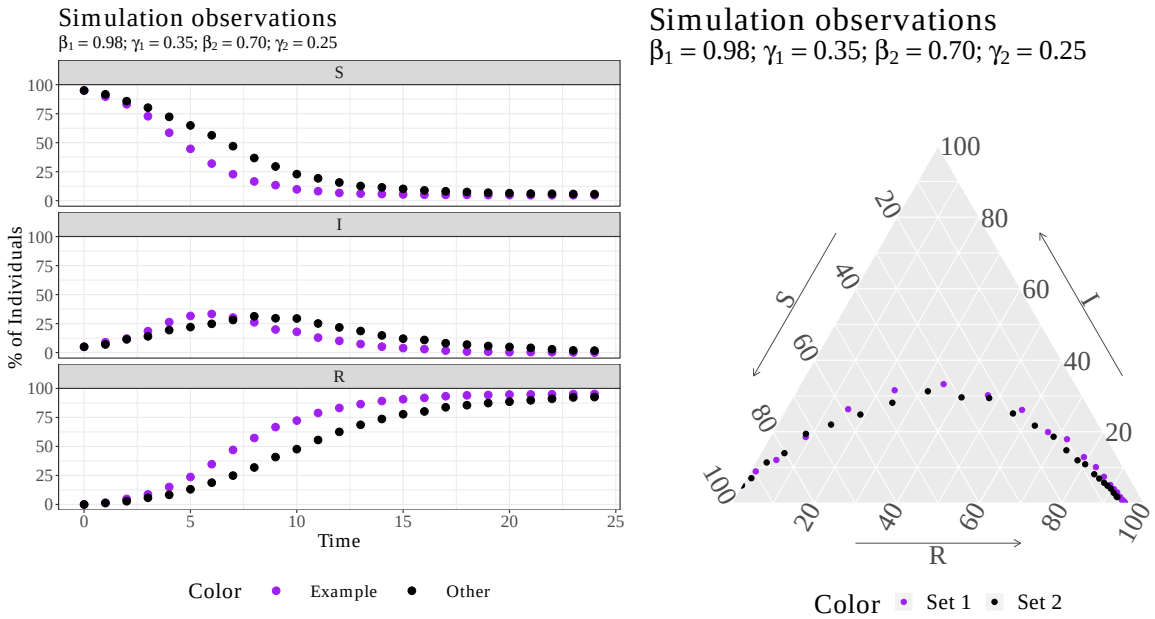

Figure 4.5: Observed SIR data simulated from the model in Eq. (2.2) with $\beta_1 = 0.98$ and $\gamma_1 = 0.35$ and a second set with $\beta_2 = 0.70$ and $\gamma_2 = 0.25$. Left: % of individuals in state vs. time. Right: ternary plot of % in S, I, and R states. The black points are from set 1 and the purple points from set 2.

In fact, the loss of the time dimension in the ternary plot can even be seen as an advantage, in some aspects. For example, viewing the % of individuals in states vs. time is dependent on the scale for time which can effect how "serious" the outbreak looks at a glance. In this regard, the ternary plot allows for a standard way to view the severity of the infection. Second of all, we can identify outbreaks that have a similar reproduction number $\mathcal{R}_0 = \frac{\beta}{\gamma}$ provided the initial conditions are similar and on average, the observations are drawn from a SIR model with $K^* = 3$ states (i.e. the individuals act homogeneously).

For example, let $S(0) = 950$, $I(0) = 50$, and $R(0) = 0$. We simulate a set of observed data from the Binomial movement model in Eq. (2.2) with $\beta_1 = 0.84$ and $\gamma_1 = 0.30$ for times $t = 0, \ldots, T$ and a second set of observed data from the Binomial movement model in Eq. (2.2) with $\beta_2 = 0.70$ and $\gamma_2 = 0.25$ for times $t = 0, \ldots, T$. Note that for both sets of observed data, $\mathcal{R}_0 = 2.80$. We plot the traditional plot (left) and ternary plot (right) in Figure 4.5. Looking at the traditional view, there is no way to determine if the $\mathcal{R}_0$ values for the two sets of observations are similar from looking at the observations alone. However, in the ternary plot, the observations are almost superimposed on one another and so we know from the plot alone that the outbreaks have similar values of $\mathcal{R}_0$. The ternary plot allows us to identify outbreaks with similar $\mathcal{R}_0$ values regardless of the value of β and γ.

Moreover, using the ternary plot we can visualize our estimate of the S, I, and R states along with point-wise confidence regions. For example, assume we have S, I, and R values from the Binomial movement

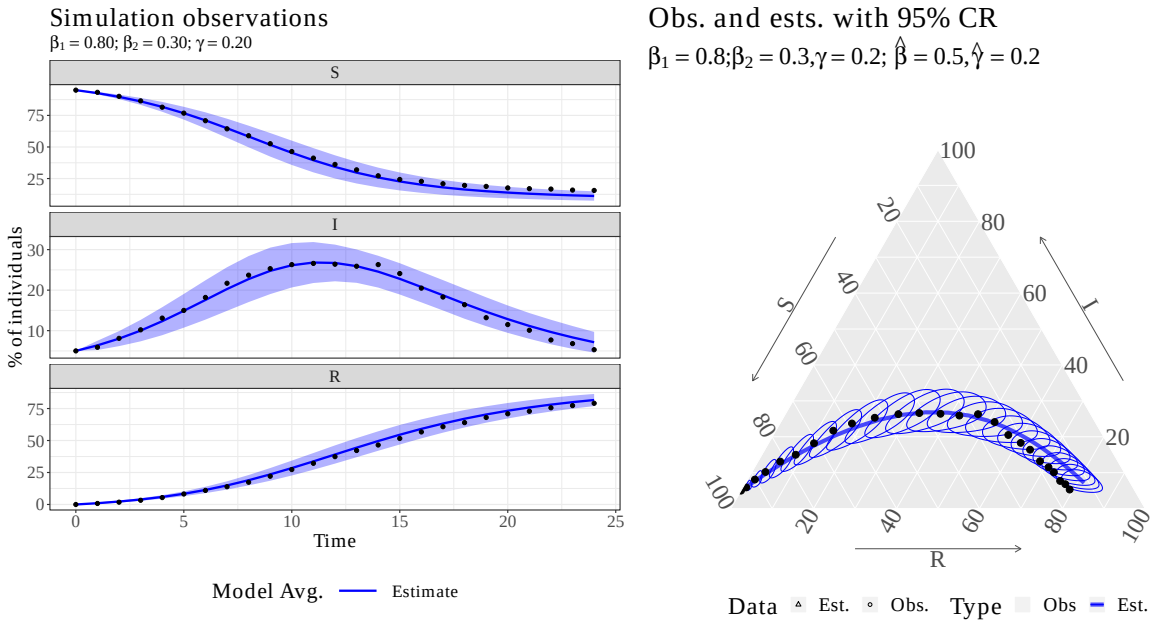

Figure 4.6: Observed SIR data simulated from a S^2IR Binomial movement model with $\beta_1 = 0.8$, $\beta_2 = 0.30$, and $\gamma = 0.20$. Our estimate of the model is $\hat{\beta}_1 = \hat{\beta}_2 = 0.5$ and $\hat{\gamma} = 0.2$. Left: average % of individuals in state vs. time as the line and the ribbon is the 95% pointwise marginal confidence intervals. Right: average % in S, I, and R states and 95% pointwise confidence regions.

model with $D(t)$ defined in Eq. (3.5). This is the model with two groups of susceptibles (e.g. males and females) who have different rates of infection β_1 and β_2, respectively and have the same rate of recovery γ.

Let $\beta_1 = 0.8$, $\beta_2 = 0.30$, and $\gamma = 0.20$. Let our estimate of the model be $\hat{\beta}_1 = \hat{\beta}_2 = 0.5$ and $\hat{\gamma} = 0.2$. Our observations and estimates are plotted in the traditional view (left) and the ternary view (right) in Figure 4.6. Generally, in both views we see that most of the points are contained the confidence intervals/regions. However, we can only see joint confidence regions in the ternary plot. This allows us to see that we begin to systematically overestimate the number of infectious and the number of recovered in later stages of the epidemic. Although we can see this systematic overestimation in the traditional view as well, it is more apparent when we can view all 3 dimensions with a single point.

The ternary plot, as implied by its name, is limited to displaying three dimensions. However, we are still able to partition the individuals into M groups and have a maximum of $3M$ total states in our stochastic SIR-CM or SIR-AM. As such, we can still incorporate heterogeneity of individuals into our SIR model and be able to assess the aggregate SIR totals using the ternary plot.

In our above example we looked at two susceptible groups of individuals with different infection parameters, β_1 and β_2. Opposed to the log-linear plot, we can still assess the fit of our model for multiple groups of S, I, or R states. In Figure 4.7 we plot two separate ternary plots for our two groups of susceptibles along with our estimates for each of the groups. In this view, it is clear we do not fit the data well as we are

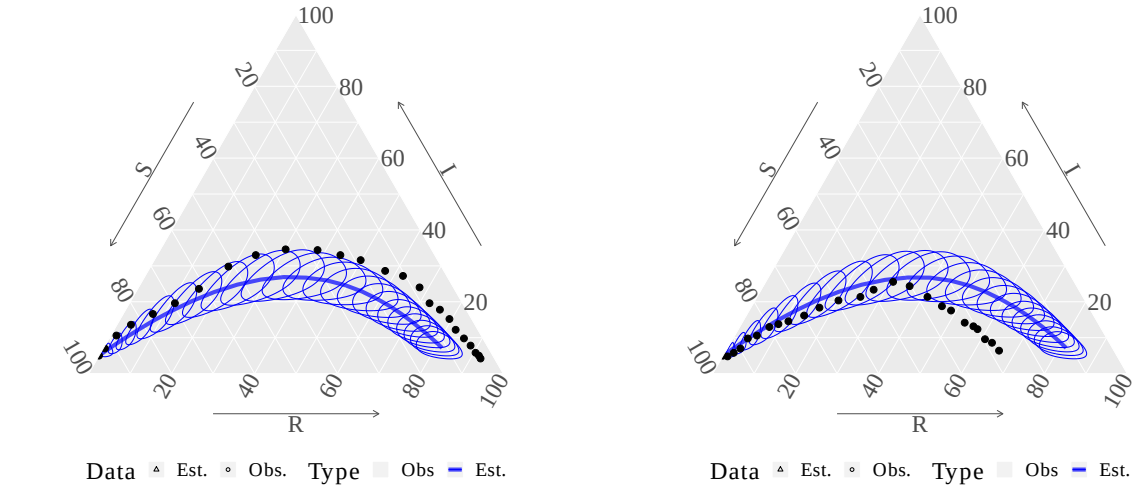

Figure 4.7: Observed SIR data simulated from a S^2IR Binomial movement model with $\beta_1 = 0.8$, $\beta_2 = 0.30$, and $\gamma = 0.20$. Our estimate of the model is $\hat{\beta}_1 = \hat{\beta}_2 = 0.5$ and $\hat{\gamma} = 0.2$. We plot the average % in S, I, and R states and 95% pointwise confidence regions for each of the two groups using ternary plots.

underestimating the number of infectious for the first group and overestimating the number of infectious for the second group.

Additionally, this idea may be extended to the SEIR disease-states using a three dimensional plot where the points are limited to a space within a tetrahedron where each side is of length 1. We will pursue this idea in the future.

In summary, the ternary plot allows us to view SIR epidemics in a standard way which allows us to more easily compare outbreaks that have different time scales. Moreover, important values such peak of infections tend to be emphasized in the ternary plot because the three dimensions are viewed as a single point. Additionally, joint confidence regions can be plotted on a ternary plot which cannot be done with the traditional view.

4.2 A statistical investigation for SI disease-level states

To introduce this concept, we begin with a small example. Unlike Chapter 2, we use an even simpler example than using SIR disease-level states. We study an example with only SI disease-level states. For these disease-level states, individuals either belong to the susceptible (S) or infectious (I) states at a given time. Once a susceptible individual becomes infectious, he will remain infectious. Here, we will use 0 to denote the S state and 1 to denote the I state. Like in Leventhal et al. (2013), models using SI disease-level states are

sometimes used to model diseases such as HIV, since there is no recovery. Using only two disease-level states, however, is often deemed to be too simple to model real situations. The "canonical" deterministic SI model is used mostly for its nice mathematical properties, at least with respect to the original deterministic SI-CM continuous time differential equation, which has a closed form solution (see Daley et al. (2001)),

$$\begin{aligned} \frac{dS}{dt} &= -\rho S\,(N-S) \\ \implies S(t) &= N - \frac{(N-S(0))N}{(N-S(0)) + (N-S(0))e^{-\rho N t}}. \end{aligned}$$

The model we study is generated from the following SI-AM for a fixed population of size N, T evenly spaced time points, $I(0)$ initially infectious individuals, and probability of infection per contact with an infectious individual ρ, and $J_{t,n}$ the number of infectious contacts agent n has at time t, and $\sigma(A_0)$, a random permutation of the vector A_0. Let $W_{t,n}$ denote a random Bernoulli variable and the agent update be given by,

$$\begin{aligned} W_{t,n} &\sim \text{Bernoulli}\left(1-(1-\rho)^{J_{t-1,n}}\right) \\ A_{t,n}|A_{t-1} &= \begin{cases} W_{t,n} & \text{if } A_{t-1,n}=0 \\ 1 & \text{if } A_{t-1,n}=1 \end{cases} \end{aligned} \tag{4.4}$$

In contrast to the AMs presented in Chapter 2- 3, we randomly assign, with equal probability, one agent to be initially infectious, which is why we require the random permutation of A_0.

The probability argument for the Bernoulli random variable $W_{t,n}$ can be explained by the fact that it is the probability of receiving the infection from at least one of the infectious contacts. This probability of transition shown in Eq. (4.4) is similar to the original Reed-Frost Chain Binomial presented in Abbey (1952).

In words, in this AM one agent is randomly assigned to be the initial infector and the remaining agents are assigned to be susceptible. A susceptible agent has an equal probability ρ of receiving the infection from an infectious contact between time step $t-1$ and time t. Thus, the probability of a susceptible agent n becoming infectious from time $t-1$ to t is dependent on the number of infectious contacts of agent n at time $t-1$, $J_{t-1,n}$. Infectious agents will remain infectious for the duration of the epidemic.

We run simulations from the AM in Eq. (4.4), and the results of the simulations are shown in Figure 4.8 (left) for $N=1000$, $\rho=0.003$, $T=50$, $I(0)=1$ and $J_{t,n}=I_t$ for all n. In this plot, we display the average percent of susceptible (blue) and infectious (red) at each time and a 95% CI. Note that the CIs are smaller for both the percent of susceptibles and the percent of infectious at both the beginning and the end of the outbreak.

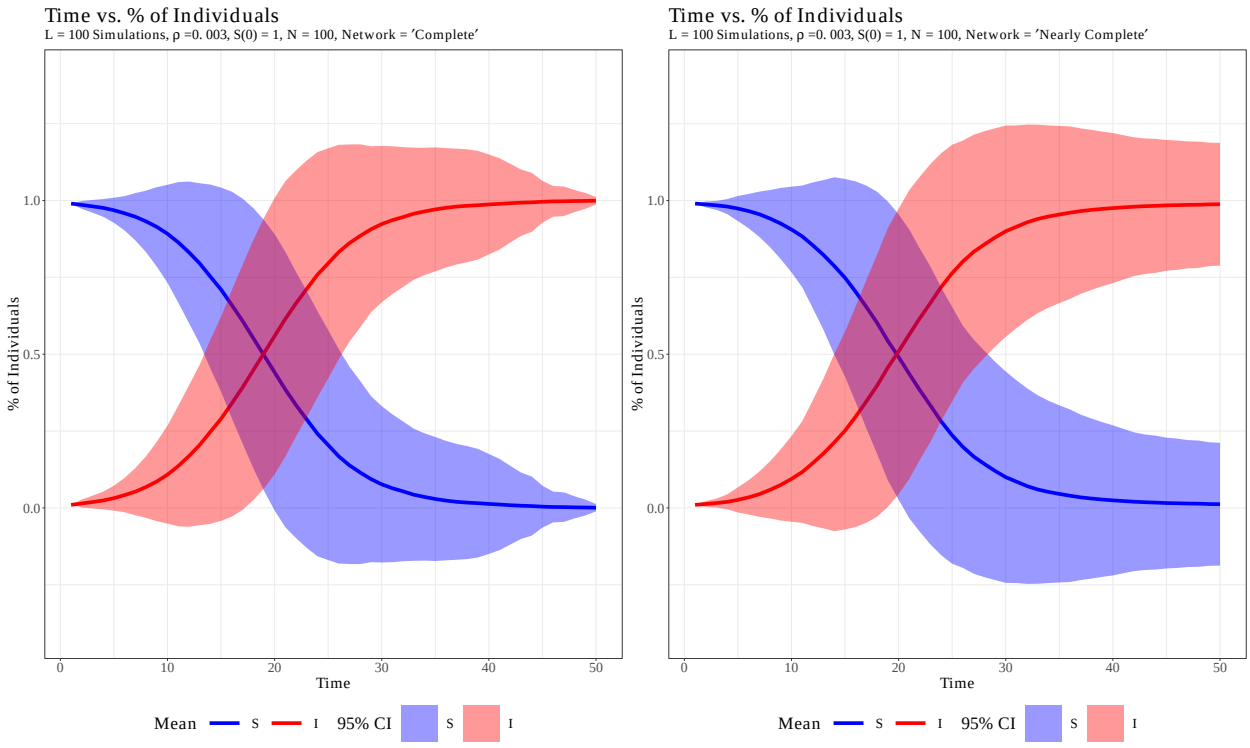

Figure 4.8: Results from simulations of Eq. (4.4) with $N = 100$, $\rho = 0.003$, $T = 50$, and $I(0) = 1$. The red horizontal line corresponds to the 95% coverage line, which is the amount of coverage we expect given our 95% prediction intervals. The red vertical line corresponds to the value 0.03, which is the value of the true recovery rate, γ, which is a point of equilibria in the K&M deterministic SIR equations.

The scenario of the number of infectious contacts of each agent being the total number of infectious agents $J_{t-1,n} = I_{t-1}$ implies homogeneous interaction of the susceptible and infectious agents. Because of this homogeneous interaction of agents, we can then write an equivalent stochastic SI-CM with $K = 2$ total states that has equivalent distribution to the AM in Eq. (4.4) in terms of the number of individuals in each state, namely the SI-CM with

$$\begin{aligned} Z_t|S_{t-1} &\sim \text{Binomial}\left(S_{t-1}, 1-(1-\rho)^{N-S_{t-1}}\right) \\ S_t|S_{t-1} &= S(t-1) - Z_t. \end{aligned}$$

In general, if the susceptible agents interact homogeneously with the infectious agents, then it is straight forward to find our equivalent stochastic CM. The question is whether we can just as easily find our equivalent stochastic CM when the agents begin to interact heterogeneously.

Take, for example, the agent contact structure in Figure 4.9 (left), where each vertex represents an agent and an edge between two agents represents a contact between the pair. In this contact structure, each agent contacts every other agent at time t for all t. More specifically, each susceptible agent contacts each infectious agent at time t and hence has the same probability of becoming infectious. This situation corresponds to $J_{t,n} = I_t$.

Once we have this agent contact structure, we can vary heterogeneity of agent interaction by removing edges between vertices, which is equivalent to taking away contacts between pairs of agents. In Figure 4.9 (right), we have taken one of the 10 agents and removed the edges to all other agents except for one. The remaining nine agents still have complete contact among one another. We call the situation in which $N-1$ agents have complete connections and the remaining agent only has one connection the "9/1" network.

We now let $J_{t,n}$ be the number of infectious contacts agent n has at time t according to the the agent contact structure in Figure 4.9

$$J_{t,n} = \#\left\{m : A_{t-1,m} = 1 \text{ and agent } n \text{ contacts agent } m\right\}. \tag{4.5}$$

The simulations produced for Eq. (4.4) for $N = 100$, $\rho = 0.003$, $T = 50$, $I(0) = 1$ and $J_{t,n}$ from Eq. (4.5) are shown in Figure 4.8 (right). In the figure, we see although the mean curves for the two states are similar to those of the complete network, the CIs are different, especially close to the beginning and the end of the epidemic.

In summary, we have described two SI-AMs that differ only by their underlying agent contact structure which determines whether the interactions of agents are homogeneous or not. The complete model has homogeneous interaction among agents whereas the nearly complete model has some heterogeneous

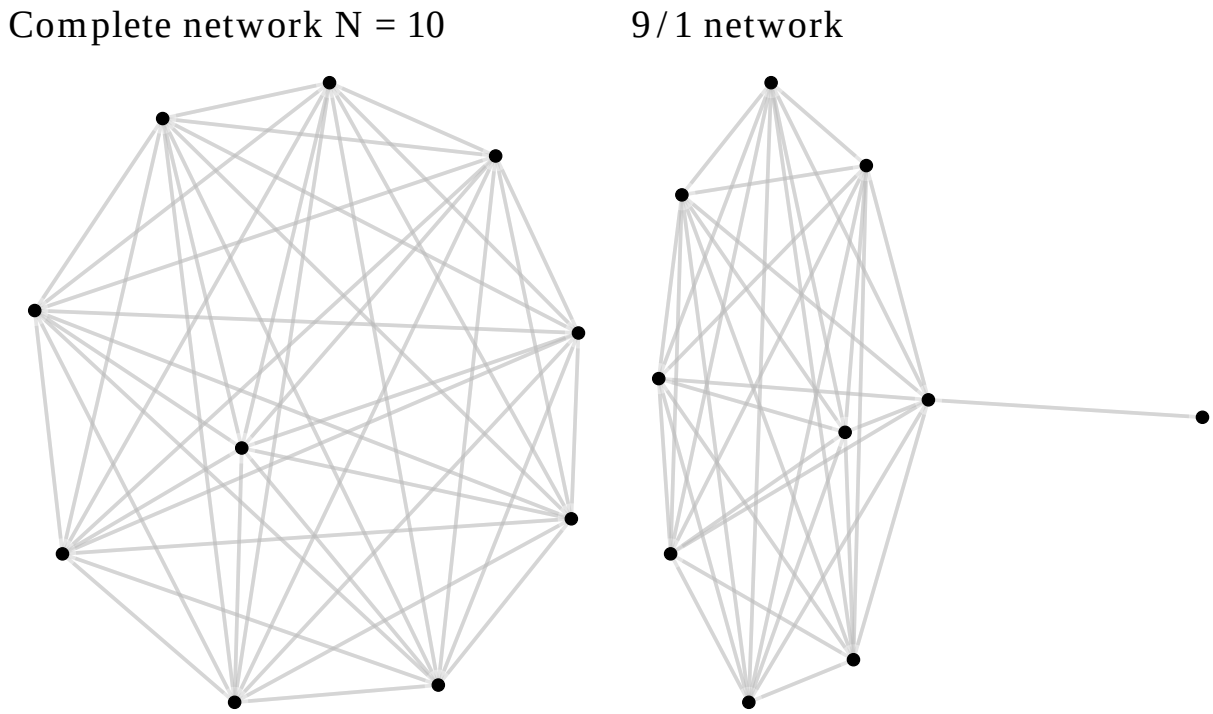

Figure 4.9: Example of a complete graph (left) and a "9/1" graph for $N = 10$ agents.

interaction of agents because of two agents (the one with only one contact and the agent is connected to every single agent).

In Figure 4.8, we see that the average number of individuals in each simulation is about the same for both the complete and 99/1 model (since there are now a total of $N = 100$ agents). The CIs for the number of individuals in each state is smaller for the complete network than the 99/1 network, notably when T is close to 50. Because these figures are so similar to one another, it makes sense to ask the question does $d(S^C, S^{99/1}) < \epsilon$ for some threshold ϵ, d is some distance function, S^C is the number of susceptibles at each time t from the complete network, and $S^{99/1}$ is the number of susceptibles at each time t from the 99/1 network. In other words, are the the distributions of the number of susceptibles in each model *similar enough* to one another? If we determine that the distributions are, indeed, close enough then we can model the 99/1 agent contact structure with $K^* = 2$ states as opposed to $K^* = 6$ (2 SI states for the 98 agents who are not connected to agent, 2 SI states for agent 2 who is connected to agent 1, and 2 SI states for agent 1 who is only connected to agent 2).

Example 4.2.1. For example, in the 9/1 in Figure 4.9 (right), we may think that if the initial infectious agent is equally likely to be any agent then the fact that the network is not complete will not influence the spread of the outbreak very much. We can quantify this with a statistical investigation. Let C be

the superscript for the complete network and 99/1 be the superscript for the 99/1 network. We want to investigate $d(T(S^C), T(S^{99/1})) < \epsilon$, where $I_0 = 1$ and the initial infectious agent is chosen uniformly at random from the set of agents and $T(S^k)$ is a statistic of S^k. This means we need some similarity or distance function to compare the two distributions. More importantly, we need to know how to interpret the magnitude of the distance between the two distributions. In this example, we will use the 5th the 95th percentiles of our estimate of ρ $(p_{2.5}, p_{97.5})$ and $\epsilon = .0001$. To be clear, we know the distributions are not the same and that is not the question we are trying to answer. We, instead, want to use the distance as a measure of *how much* the distributions differ.

The process for this investigation is to fix ρ and the agent contact structures for the complete and 99/1. Then we

1. Generate L data sets, assuming the first contact structure i
2. Estimate $\hat{\rho}^i_\ell$ for $\ell = 1, \ldots, L$ assuming homogeneous interaction of agents (e.g. a SI-CM with $K = 2$ states)
3. Set $\hat{F}^i$ as the empirical distribution of $\hat{\rho}^i_\ell$
4. Repeat for the other contact structure j

We then use the Euclidean distance between the two sets of percentiles for our estimate of ρ. We demonstrate this with the estimated $\hat{\rho}$ values from simulations of the complete and nearly complete graph in Figure 4.10 with $L = 100$ simulations. For the most part, the estimates of ρ look to have a similar distribution. For the nearly complete network, we see an observation of $\hat{\rho} = 0$ which can be attributed to when the initial infectious agent was the agent who was connected to only one other agent.

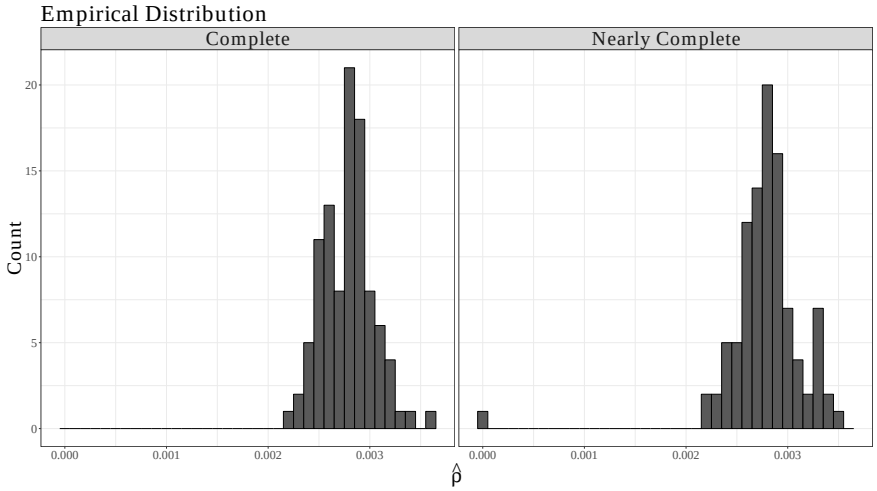

Figure 4.10: Estimates of $\hat{\rho}$ from simulating an SI outbreak from a complete agent network $\mathcal{G}^C$ (left) and a nearly complete agent network $\mathcal{G}^{NC}$ (right). In both simulations, the initial infectious agent is chosen uniformly at random.

The distance is $d = 6 \times 10^{-5} < \epsilon$ and so we may conclude that the two models are similar enough for our purposes. In turn, we could model the population with $K^* = 2$ states. That is, we determine that S^C and $S^{99/1}$ are similar enough to one another that we accept the small differences in distribution for simplicity of the model with fewer parameters.

The benefits from being able to estimate a nearly complete agent interaction structure with a complete interaction structure include (1) reducing variance in our estimates (which we explore more in Ch. 9), (2) computational and memory speed-ups from treating agents as indistinguishable from one another, and (3) model simplification.

That said, there are some difficulties in this process, which include (1) possibly having to estimate θ, the vector of disease parameters; (2) selecting an adequate CM; and (3) having to choose a meaningful distance d, a statistic T, and threshold ϵ. In the next chapter, we focus on issues (1) estimating θ and (2) selecting an adequate CM. With regards to issue (3), the SI disease-level states represent a contrived example as it only has two disease-level states, S and I, and subsequently one disease parameter, ρ, and so it easier to select a meaningful distance, statistic, and threshold. Additionally, in the example we provided the true distributions of S^C and $S^{99/1}$ are *not* equivalent. In this case, we would likely accept the slight difference in distribution in order to have a simpler and more efficient model. Future work in this area will focus on quantifying the difference between the empirical distributions and what constitutes an acceptable difference.

4.3 Chapter summary

In this chapter we present visual diagnostics and a statistical investigation to aid in selection of a stochastic CM-AM pair with K total states when the fixed disease-level states are either SIR or SI. As ultimately we want to infer information about a disease, it is important to choose a model that fits the observed data well. We examine ways to choose K^*, the minimal number of states needed to adequately model a CM/AM pair which include examining the interaction structure of agents and visual diagnostics.

To improve model selection, we present two novel ways to visualize observed and estimated data for SIR disease-level states. The first is a log-linear formulation of the model, which is derived from using the result of Harko et al. (2014) that reduces the two-dimensional continuous time differential equations for the deterministic SIR model into a one-dimensional equation. This, in turn, indicates how to transform our raw SIR data into the form of simple linear regression through the origin where $\mathcal{R}_0$, the reproduction number, corresponds to best-fit line. When using weighted linear regression where the weights are the inverse of the estimated plug-in variance from Eq. (2.4), we can generate prediction intervals that cover 95% of the data and is robust to our choice of β. As such, the log-linear plot and 95% prediction interval can be used as a check to see whether a stochastic SIR model with $K^* = 3$ states is a good fit for data.

The second visualization we present is a ternary plot for SIR disease-level states. This plot can be used to assess the fit for $K^* \leq 3M$ total states where M is the number of groups in which we partition the individuals. We present an example in Figure 4.7. Not only does the ternary plot allow us to visualize all three dimensions of the data (SIR) at once, we can also plot 95% pointwise CRs and color observations by time. As such, the ternary plot provides a very flexible way to visualize SIR data, regardless of homogeneous interaction of agents.

The next method we present to aid in selection of the minimal number of states is specific to the SI disease-level states. This method answers the question of whether a population is homogeneous "enough", in the sense that a stochastic SI-CM with $K^* = 2$ total states can be used to model the population. We present a method to use to compare estimated ρ values from a non-homogeneous interaction of agents model to one with homogeneous interaction. Our small example shows a case where a population with one outlier individual that does not interact homogeneously with the rest of the population can be adequately modeled with a homogeneous model. The upshot of this method is that the complexity of a model can be reduced greatly if we assume homogeneous interaction, which in turn can increase the interpretability of the model and also be involved in practical speed-ups as simulation time is greatly reduced. The method we present is limited to the SI disease-level states due to the difficulty in selecting a distance, a statistic, and threshold to compare our models to one another.

In the next set of chapters we will use some of the methods presented in this chapter along with more traditional model selection techniques (such as MSE and AIC) to analyze two historical outbreaks: measles in Hagelloch, Germany (1861-1862), and Ebola in Western District, Sierra Leone (2014-2017).

Chapter 5

Measles: model selection

In Chapters 2 and 3 we discussed conditions needed to have equivalent stochastic CM-AM pairs, and in this chapter we apply that theory and methodology to real data. More specifically, we examine data from a measles outbreak in 1861-1862 in Hagelloch, Germany. Our goals in this analysis are to:

1. Identify groups of individuals that behave differently from one another
2. Determine the minimal number of states K^* needed to generate a stochastic CM-AM pair
3. Estimate parameters for our CM-AM pair
4. Estimate $\mathcal{R}_0$ and compare to other measles outbreaks
5. Use our paired AM to implement hypothetical scenarios and prevention policies.

In this chapter, we will focus on issues (1)-(4) and in Chapters 6-7, we will examine issue (5) in depth.

After introducing the data set, we begin our search for adequate CM-AM pairs. To do so, we first fix disease level-states. Here, we decide on the SIR disease-level states, for reasons discussed below. Once the disease-level states are fixed, we can focus on finding a minimal number of states, K^*, and ultimately find the corresponding states associated with such a model and estimate disease parameter estimates to use in our paired AM. To do so, we examine a variety of models and modeling assumptions. To assess our models, we look at both quantitative and qualitative methods including mean square error, (MSE) the Akaike information criterion (AIC), and three diagnostic plots, two of which are novel. Finally, we use our best estimates to address the question of what the value $\mathcal{R}_0$ is in this outbreak.

This chapter proceeds as follows. In Section 5.1, we introduce the Hagelloch data set and conduct exploratory data analysis (EDA). In Section 5.2, we provide the reasoning behind our models, the methodology used to fit them, and our resulting assessment of the best models. Finally, in Section 5.3 we summarize the findings in this chapter.

5.1 Data and EDA

Before we address issues (1)-(4), we introduce the data set. Despite being over 150 years old, this data set is still relevant today because (1) measles is still relevant (Stobbe, 2019; Balser, 2019), (2) the data set provides a good testing ground for methodology due to feature rich data and the small, isolated nature of the village, and (3) many features of the data set are very applicable to modern models including spatio-temporal and network analysis Liu et al. (2015a); Lessler et al. (2016); Groendyke and Welch (2018). We give a high level overview of the data along with some exploratory data analysis (EDA), and more EDA is available for interested readers in Appendix A.

The Hagelloch data was initially collected by Pfeilsticker (1863) and further analyzed by Oesterle (1992). The data set follows the course of a measles epidemic in Hagelloch, Germany from October 30, 1861 (Day 0) until January 24, 1862, covering a period of 87 days. Figure 5.1 shows a satellite image of current day Hagelloch.

Figure 5.1: Current day satellite image of Hagelloch, Germany

Along with mumps, rubella, and varicella, measles is a highly infectious childhood disease. Symptoms of of the disease include high fever, cough, runny nose, and red, watery eyes. Two to three days after initial symptoms, tiny white spots may be found in the mouth. Three to five days after the symptoms begin, a rash appears on the body. A high fever (104 degrees F or more) may also be observed. Finally, the rash and fever resolve after a few days (Centers for Disease Control and Prevention, 2018).

Table 5.1: Subset of data from the `surveillance` package in `R`. The time of initial infectiousness (Time I) and initial recovery (Time R) are imputed in Salmon et al. (2016) from the recorded symptom appearances.

ID	Surname	Age	Sex	Time I	Time R	Infector ID
1	Mueller	7	Female	22	30	45
2	Mueller	6	Female	24	32	45
3	Mueller	4	Female	29	37	172
45	Goehring	7	Male	12	18	184
184	NA	13	NA	0	11	NA

Measles is transferred from person to person through contaminated air or an infected surface. Centers for Disease Control and Prevention (2018) reports that a person is infectious four days before and after the appearance of the rash. Measles is perhaps the most contagious person-disease on the planet, with a reproductive number estimated of around $\mathcal{R}_0 = 19$ (Anderson and May, 1992), which means that when an infectious person is introduced to a fully susceptible population, she will infect on average 19 others. In fact, a seven year-old boy in the Hagelloch data set purportedly infected 30 other individuals. However, more recent estimates of $\mathcal{R}_0$ for measles are closer to 6-7 (Getz et al., 2016).

This Hagelloch data set, available from the `surveillance` package in `R`, includes 188 cases of measles out of 568 total inhabitants (Salmon et al., 2016). Neal and Roberts (2004) argue that the 188 cases were the only such individuals who were susceptible to the disease due to an outbreak 14 years prior. We also will maintain that argument and therefore use $N = 188$ as the total susceptible population.

The data collected is feature rich, especially considering the fact that epidemic data is typically not publicly available due to privacy concerns. All reported cases in the data set experienced the measles rash. The data set includes sex, age, class level (preschool, 1st class, or 2nd class), household ID, and location (x, y coordinates), date of first appearance of symptoms and duration of early symptoms, date of measles rash, unique family ID, purported infector ID, time of death, and other complications such as bronchitis. A subset of the data is displayed in Table 5.1.

In many ways, this data set is ideal for testing and analyzing methodology for the spread of infectious disease. The village is small and fairly isolated with a fairly homogeneous population in terms of ethnicity and socio-economic status. A strong argument exists that the only susceptible people in the population are the 188 children who did become infected since measles typically occurs during childhood. Furthermore, we have household location, household ID, sex, age, and class as features of our children which lends itself to spatio-temporal modeling. In terms of network structure, class and household structure are features that can be used to estimate social networks. Moreover, the purported source of infection is reported for 184 out of the 188 children which allows us to analyze the spread of infection over a network.

As previously mentioned in Chapter 3, in order to examine CM-AM pairs, we need to first decide on which disease-level states to use in our models. Besides, the S and I states, it makes sense to have a R state

in which individuals can move to over time, as individuals either recover and are no longer susceptible or die from measles. The original data has date of first symptoms (prodromes) along with the appearance of the rash date. From this, Salmon et al. (2016) impute the time of infection and time of recovery for each individual based on the estimates in Neal and Roberts (2004). We use the imputed time of infection and time of recovery times, as recorded in the `surveillance` package in our analysis.

We plot the number of susceptible, infectious, and recovered children over time in Figure 5.2. In the figure, we see that the number of susceptible children is non-increasing and the number of recovered children is non-decreasing, as we would expect. However, we note that there are two local maxima for the number of infectious children, the first around November 25 and the second around December 5. In this analysis, we show that the first maximum is unlikely to be attributed to random noise. The existence of multiple maxima in terms of the number of infectious individuals makes us question whether it is appropriate to fit a stochastic SIR-CM to the data because under the the *deterministic* SIR-CM, the number of infectious necessarily has one maximum. While the stochastic SIR-CM is more flexible than its deterministic counterpart and can account for infectious curves with more than one maximum, it gives us reason to believe that the agents in this data set may either have differences in susceptibility to measles, differences in mixing with other agents, or both.

We analyze whether the spatial location of the households affects the spread of the disease. In the original data set, the x and y location of the children was plotted on a 250×250 m^2 grid. Unfortunately, it is unknown how this grid overlays with the current map of Hagelloch displayed in Figure 5.1. In Figure 5.3, we plot the household locations from the grid and color the children in the household by class. As we can see from the figure, most of the households are located on on the right side of the grid, with a few outliers on the left. Additionally, quite a few households have more than one infection recorded. The median number of total infections per household with at least one infection is three and the mean is 3.36 infections. There are three households with eight infected children.

Of the 184 infections where the alleged infectee was recorded, 90 infectors belonged to the same household as the infectee. This means that sibling-sibling disease transfer accounts for nearly half of the spread of the disease. Overall, we find household and class features account for over 90% of the spread of the disease.

5.2 Modeling, likelihood, and parameter estimation

In Section 5.1 we looked at possible driving forces of the resulting measles outbreak, and in this section we examine different models to adequately model the outbreak. The ultimate goal is to obtain parameter estimates to use in our paired stochastic CM-AM. We must first select one or more models.

We limit our models to those with the disease-level states SIR due to the nature of how measles is spread and data limitations. For example, although an exposure state where the agent is infected but not yet

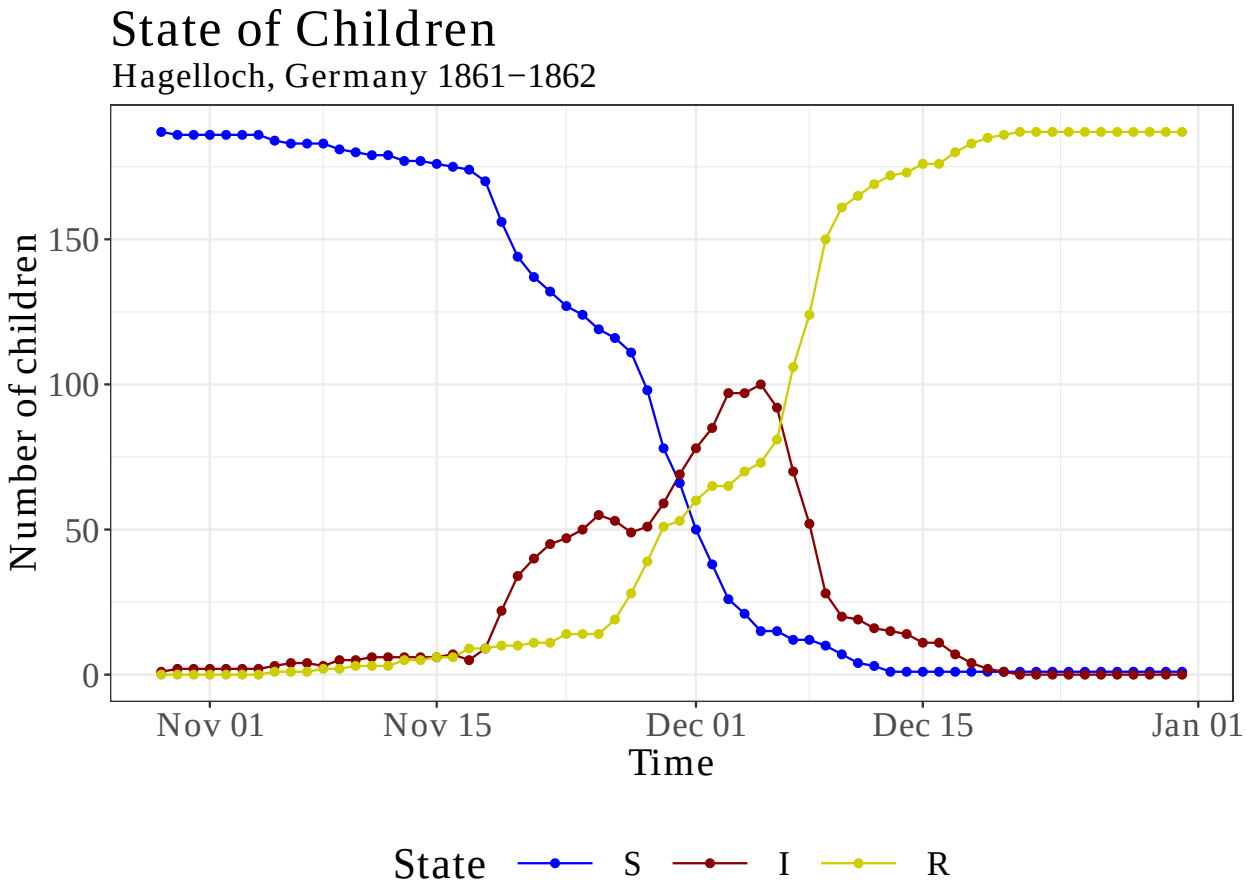

Figure 5.2: The number of susceptible, infectious, and recovered children over time.

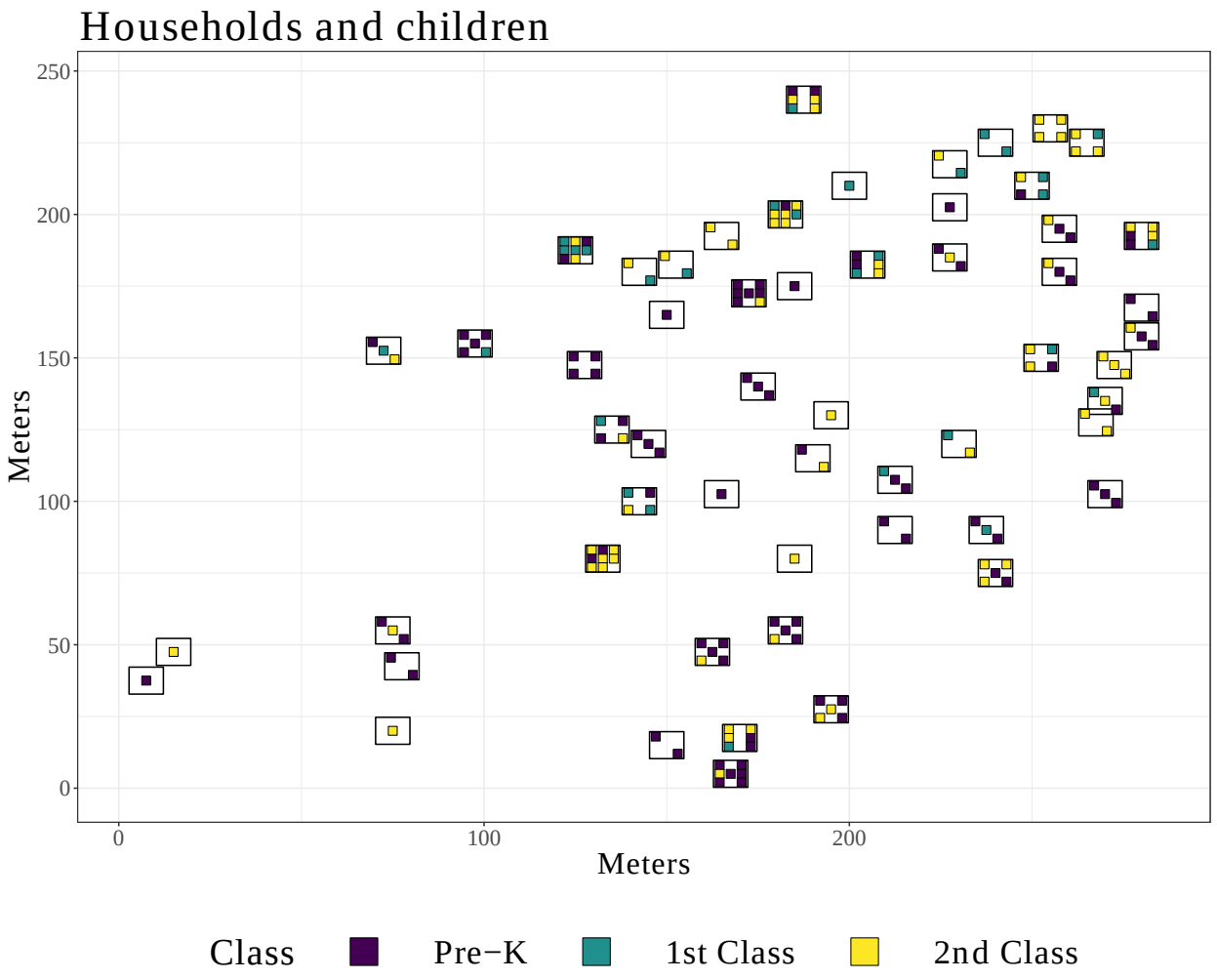

Figure 5.3: Grid of locations of infected households in Hagelloch colored by the school class of each child

infectious may be relevant, we have little to no data to assess this. For the fixed level disease-states of SIR, there are two sets of sufficient statistics we can work with depending on which modeling view we take and depending on whether we consider individuals as homogeneous within states. We call the first statistic, when the individuals are homogeneous in states, $\mathbf{X}$, and the second, when the individuals are heterogeneous or distinguishable, $\mathbf{U}$. Once we have our sufficient statistics (data), we can then calculate the likelihood of our stochastic CM, which is dependent on the total number of state K^*. Finally, to estimate our parameters, we maximize the likelihood of our models. To compare our models to one another, we look at both mean squared error (MSE) and Akaike Information Criterion (AIC) (Wasserman, 2004) as quantitative measures and three diagnostic plots as qualitative measures.

We summarize the process of how we select our models below.

1. We transform our data into two sets of summary statistics.
    - The statistic $\mathbf{X}$ is the number of susceptible, infectious, and recovered individuals at each time step and is used in models that assumes individuals are **homogeneous** and indistinguishable from one another.
    - The statistic $\mathbf{U}$ details when each individual becomes infectious and recovers and is used in models that assumes individuals are **heterogeneous** and distinguishable from one another.
2. We determine which agents behave similarly and thus can be grouped together.
    - We use basic statistical clustering methods to examine which agents behave similarly.
    - We also examine differences of agents based on time of infection.
3. We use the groupings found to identify models, estimate parameters, and compare the models, and select the best one(s).
    - We compare the models **quantitatively** with MSE, AIC, and $\mathcal{R}_0$ estimates.
    - We compare the models **qualitatively** using visual diagnostics.

5.2.1 Sufficient statistics

The first step is to organize the data into sufficient statistics. How we view the data and subsequently calculate the likelihood of the models may change depending on how the data are collected and whether we can distinguish between individuals. The first set of sufficient statistics is

$$\mathbf{x} = \{\mathbf{x}_t : (x_{t,1}, \dots x_{t,K}) \text{ for } t = 0 \dots, T\}, \tag{5.1}$$

Table 5.2: Turning the raw Hagelloch data of Table 5.1 into a sufficient statistic $\mathbf{X}$ based on the number of susceptible, infectious, and recovered at each time point. A subset is shown here.

Day	$x_{t,1}$	$x_{t,2}$	$x_{t,3}$
10	183	4	1
20	175	7	6
30	116	53	19
40	15	92	81
50	1	11	176

which is the number of agents in each state k for $k = 1, \ldots, K$ for times $t = 0, \ldots, T$, given there are K total states and the total population size is fixed. An example of this statistic is given in Table 5.2.

For the SIR disease-level states with $K = 3$ states, we look at $(x_{t,1}, x_{t,2}, x_{t,3})$, the number of individuals in the susceptible, infectious, and recovered states, respectively at time t. We assume that the individuals update according to the stochastic CM in Eq. (2.2), which is based on Binomial draws. We are interested in the likelihood of $\theta = (\beta, \gamma)$, conditioned on the number of individuals at the previous time step,

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}) &= \prod_{t=1}^{T} P(\mathbf{X}_t = \mathbf{x}_t | \mathbf{X}_{t-1} = \mathbf{x}_{t-1}, \theta) \\ &\propto \prod_{t=1}^{T} \left[p_t(\theta)\right]^{x_{t-1,1} - x_{t,1}} \left(1 - \left[p_t(\theta)\right]\right)^{x_{t,1}} \gamma^{x_{t,3} - x_{t-1,3}} \left(1 - \gamma\right)^{x_{t-1,2} - (x_{t,3} - x_{t-1,3})} \end{aligned} \tag{5.2}$$

where the specification of $p_t(\theta)$ is based on one of two possible models. The first specification (i) $p_t(\theta) = \frac{\beta}{N} X_2(t-1; \theta)$ corresponds to the deterministic Kermack and McKendrick formulation of an SIR model. The second specification (ii) $p_t(\theta) = \left(1 - \frac{\beta}{N}\right)^{X_2(t-1;\theta)}$ is taken from the framework of chain Binomial Reed-Frost models (Abbey, 1952). If β is small, then the two expressions are approximately equivalent. Note that both these probabilities $p_t(\theta)$ are deterministic as $X_2(t, \theta)$ is the deterministic number of infectious individuals at time t as a function of θ.

If we consider the agents to be heterogeneous/distinguishable from one another then a sufficient statistic consists of an agent's initial state, the maximum time the agent was still susceptible, and the maximum time the agent was still infectious,

$$\mathbf{u} = \left\{ \mathbf{u}_n = (a_{0,n}, t^*_{1,n}, t^*_{2,n}) \text{ for } n = 1, \ldots N \right\}$$

where

$$t^*_{1,n} = \min\left\{\max_{t=1,\ldots,T}\{t : A_{t,n} = 1\}, T\right\} \text{ (maximum time susceptible)}$$
$$t^*_{2,n} = \min\left\{\max_{t=1,\ldots,T}\{t : A_{t,n} = 2\}, T\right\} \text{ (maximum time infectious).}$$

In this way, we can track the states of individuals over time, as opposed to using $\mathbf{X}$, where we only know the aggregate number of individuals in each state at a given time. The likelihood is then dependent on the initial state of the individuals where $\mathcal{I}\{\cdot\}$ is the indicator function of its arguments,

$$\begin{aligned}
\mathcal{L}(\theta; \mathbf{u}) &= \prod_{n=1}^{N} P(\mathbf{U}_n = \mathbf{u}_n) \\
&= \prod_{n=1}^{N} \Big(\mathcal{I}\{A_{0,n} = 1\}\left[P(\mathbf{U}_n = \mathbf{u}_n | t^*_{1,n} < t^*_{2,n} < T) + P(\mathbf{U}_n = \mathbf{u}_n | t^*_{1,n} < t^*_{2,n} = T) + P(\mathbf{U}_n = \mathbf{u}_n | t^*_{1,n} = T)\right] \\
&+ \mathcal{I}\{A_{0,n} = 2\}\left[P(\mathbf{U}_n = \mathbf{u}_n | t^*_{2,n} < T) + P(\mathbf{U}_n = \mathbf{u}_n | t^*_{2,n} = T)\right] \\
&+ \mathcal{I}\{A_{0,n} = 3\}\Big), \qquad (5.3)
\end{aligned}$$

with

$$\begin{aligned}
P(\mathbf{U}_n = \mathbf{u}_n | t^*_{1,n} < t^*_{2,n} < T) &= \left[\prod_{t=1}^{t^*_{1,n}}(1 - p_t(\theta))\right] \cdot p_{t^*_{1,n}+1}(\theta) \cdot (1-\gamma)^{t^*_{2,n} - t^*_{1,n} + 1} \cdot \gamma \\
P(\mathbf{U}_n = \mathbf{u}_n | t^*_{1,n} < t^*_{2,n} = T) &= \left[\prod_{t=1}^{t^*_{1,n}}(1 - p_t(\theta))\right] \cdot p_{t^*_{1,n}+1}(\theta) \cdot (1-\gamma)^{T - t^*_{1,n} + 1} \\
P(\mathbf{U}_n = \mathbf{u}_n | t^*_{1,n} = t^*_{2,n} = T) &= \left[\prod_{t=1}^{T}(1 - p_t(\theta))\right] \\
P(\mathbf{U}_n = \mathbf{u}_n | t^*_{2,n} < T) &= (1-\gamma)^{T - t^*_{2,n} + 1} \cdot \gamma \\
P(\mathbf{U}_n = \mathbf{u}_n | t^*_{2,n} = T) &= (1-\gamma)^{T - t^*_{2,n} + 1}, \qquad (5.4)
\end{aligned}$$

where we assume $t^*_{1,n}$ and $t^*_{2,m}$ are independent of one another when $n \neq m$, in order to have a computationally tractable model. The likelihood may seem complicated but is simply partitioned by possible states an agent may move to (or not) over the course of an epidemic. An example of a subset of the $\mathbf{U}$ sufficient statistic from the Hagelloch data set is shown in Table 5.3.

Table 5.3: Turning the raw Hagelloch data of Table 5.1 into a sufficient statistic **U** based on the initial state, infection date, and recovery date of each individual. A subset is shown here.

ID	$a_{0,n}$	$t^*_{1,n}$	$t^*_{2,n}$
1	0	22	30
2	0	24	32
3	0	29	37
45	0	12	18
184	1	0	11

Table 5.4: Result of $K^* = 3$ SIR model fits to the Hagelloch data.

Model No.	$p_t(\theta)$	Suff. Stat	Fit	$\hat{\beta}$	$\hat{\beta} - 2SE(\hat{\beta})$	$\hat{\beta} + 2SE(\hat{\beta})$	$\hat{\gamma}$	$\hat{\gamma} - 2SE(\hat{\gamma})$	$\hat{\gamma} + 2SE(\hat{\gamma})$	Log Like.	MSE
1	KM	X	LL	0.279	0.263	0.295	0.100	0.086	0.113	-1247.767	1190.633
2	RF	X	LL	0.279	0.263	0.295	0.100	0.086	0.114	-1248.237	1203.439
3	KM	U	MSE	0.278	0.276	0.281	0.089	0.086	0.091	NA	1107.003
4	RF	U	MSE	0.275	0.273	0.278	0.090	0.087	0.092	NA	1116.495
5	KM	U	LL	0.279	0.263	0.295	0.099	0.086	0.113	-1247.872	1190.408
6	RF	U	LL	0.279	0.263	0.295	0.100	0.086	0.114	-1248.343	1203.155

5.2.2 Model fitting and parameter estimation – finding K^*

In order to find K^*, the minimum number of states to adequately model the Hagelloch epidemic, assuming the $M = 3$ SIR disease-level states, we fit a number of models to the observed data.

We first determine whether the minimum number of states of $K^* = 3$ (one S, I, and R state for the entire population) is adequate. If so, we can assume that the population mixes homogeneously. We fit six different models. The models are fit using one of maximizing the likelihood in Eq. (5.2) with sufficient statistic **X**, maximizing the likelihood of Eq. (5.3) with sufficient statistic **U**, or minimizing the mean square error where the "truth" is taken to be the deterministic K&M difference equations as a function of β and γ. For each of these fit methods, we use both the probabilities $p_t(\theta)$ of becoming infectious, the Kermack & McKendrick (KM) and Reed-Frost (RF) formulation, respectively. For the sample error of the parameters, we use a numerical estimate of the variance using the second derivative of the log likelihood for (β, γ), i.e. the observed Fisher Information. For the MSE, we use the inverse Hessian of the optimized parameters as an estimate of the covariance matrix. The results of model fitting are displayed in Table 5.4, and the estimated fits alongside the observed values are shown in Figure 5.4.

The models in Table 5.4 yield $\hat{\beta} \approx 0.28$. The log likelihood fit models yield $\hat{\gamma} \approx 0.10$ whereas the MSE yields $\hat{\gamma} \approx 0.09$. The log likelihood for the the four methods using that method to fit the data ranges between -1249 and -1247. The MSE for the methods fit by MSE have a smaller value than the log likelihood methods by an order of 10, and the MSE methods unsurprisingly having the smallest MSE. However, the root MSE only differs by approximately one person.

The fitted models plotted alongside the observations in Figure 5.4 are very similar to one another although models 3 and 4 (the MSE fit methods) seem to have more striking differences compared to log likelihood fit

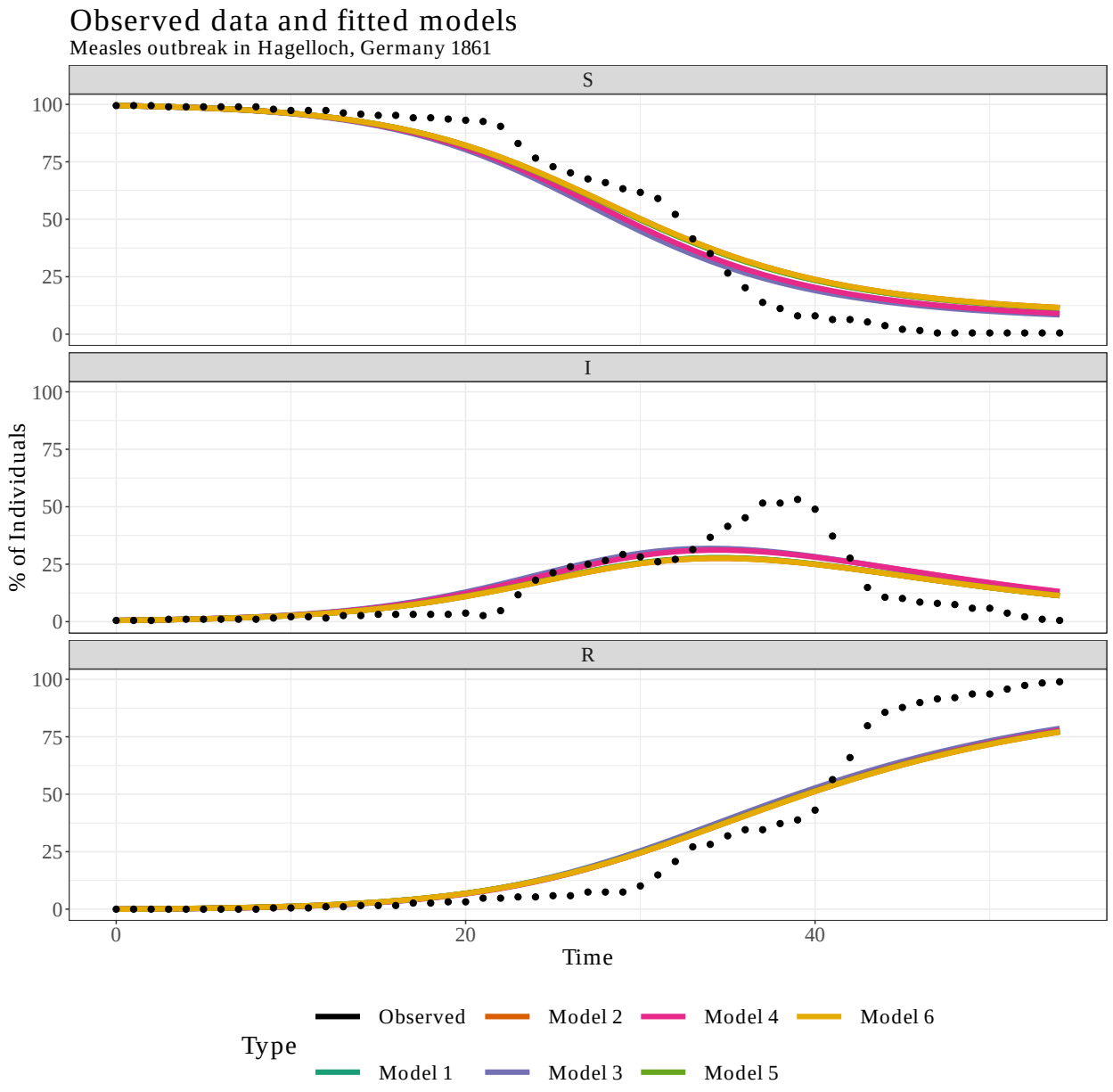

Figure 5.4: Model fits for $K^* = 3$ to the Hagelloch measles data.

models. We see none of the models fit the observed data particularly well (see Figure 5.4). This is especially apparent for the infectious state and the number of recovered individuals at the end of the outbreak.

We can also view our results using ternary plot as described in Chapter 4. We display the results of the same methods used in Figure 5.4 in Figure 5.5. In this coordinate system, the two peaks in the observed data infectious curve in Figure 5.4 become even more apparent. It is clear that the SIR model estimates do not fit the data well, especially after day 30. We also see a more apparent difference between the set of models 3 and 4 (the MSE fit methods) and the set of models 1, 2, 5, and 6 (the log likelihood fit methods).

Finally, we can view our estimates in the log linear formulation, according to Eq. (4.2). In this visualization, we would expect, on average, that our transformed variables would form a straight line whose slope corresponds to $\mathcal{R}_0$, the reproduction number. The data and estimates are plotted in Figure 5.6. The model estimates form straight lines, which is what we would expect given the structure of the models because the expected value of each state in the models is equal to the number of individuals in each state in the deterministic K&M difference equations. None of the models seem to fit the observed data particularly well. In this view, we can see that the $\hat{\mathcal{R}}_0$ estimates for the log linear fits are close to 2.80 (95% CI [2.73, 2.88] for Model 1) and for the MSE estimate are closer to 3.12 (95% CI [2.94, 3.26] for Model 7). Notably, both of these estimates for $\hat{\mathcal{R}}_0$ are much less than 19, which was the estimate for measles but is still fairly large relative to other disease $\hat{\mathcal{R}}_0$ estimates for other diseases such as influenza and Ebola as reported in Anderson and May (1992).

In fact, throughout this entire chapter we will see that our largest estimate of the reproduction number $\hat{\mathcal{R}}_0$ is about 5 and so the reader may wonder why there is such a large discrepancy between this and the Anderson and May estimate. One possible explanation is that the susceptible population is equal to a subset of the village population rather than the entire population. Another possible reason is that in the raw data, one child is purported to have infected 26 classmates out of 28 classmates, which is an influential event. Finally, Getz et al. (2016) estimate $\mathcal{R}_0$ =5-6 in a more recent epidemic outbreak in India.

We can also model the observed data using the weighted linear regression through the origin estimate, according to the equation in Eq. (4.2) where we weight each point by the inverse of its estimated variance.

The linear regression estimate is not a generative model as R_t and S_t are both random variables. That said, we can still use it learn about the outbreak. This weighted linear regression model is plotted in Figure 5.7. Compared to the model estimates in Figure 5.6, the weighted linear regression line seems to fit the data much better and the 95% point-wise prediction interval covers the points well. Intuitively, it makes sense that the prediction interval increases in width over time.

With the weighted linear regression model, we estimate $\hat{\mathcal{R}}_0$ as 4.94 (95% CI: [4.68, 5.21]), which is substantially larger than the $\hat{\mathcal{R}}_0$ estimates from the models in Figure 5.6. If the model did follow the SIR formulation with $K^* = 3$, then we would expect the weighted linear regression estimate of $\hat{\mathcal{R}}_0$ and the estimates from our models to be very similar, but this is not the case here.

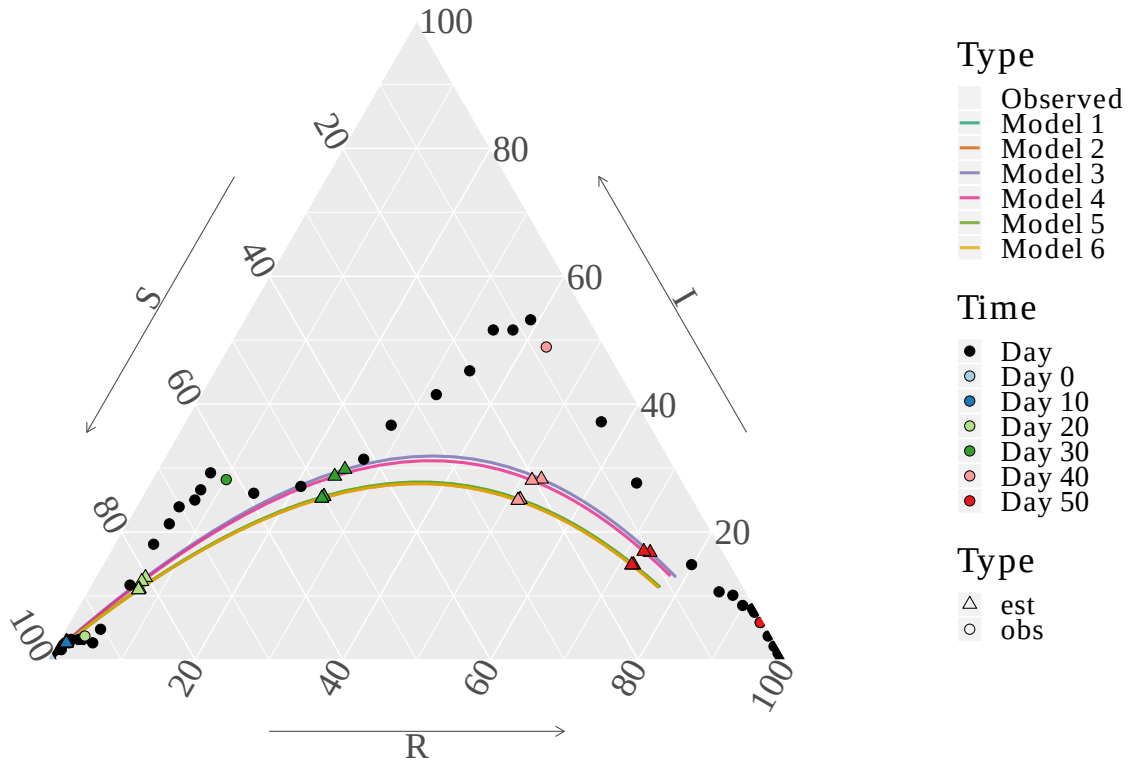

Figure 5.5: Model fits for $K^* = 3$ to the Hagelloch measles data plotted in barycentric coordinates via a ternary plot. The observed estimates are plotted as circles and the estimates are plotted as triangles. Every 10th day is filled in with a different color in order to identify points that occur at the same time. The different model estimates are plotted with different color lines.

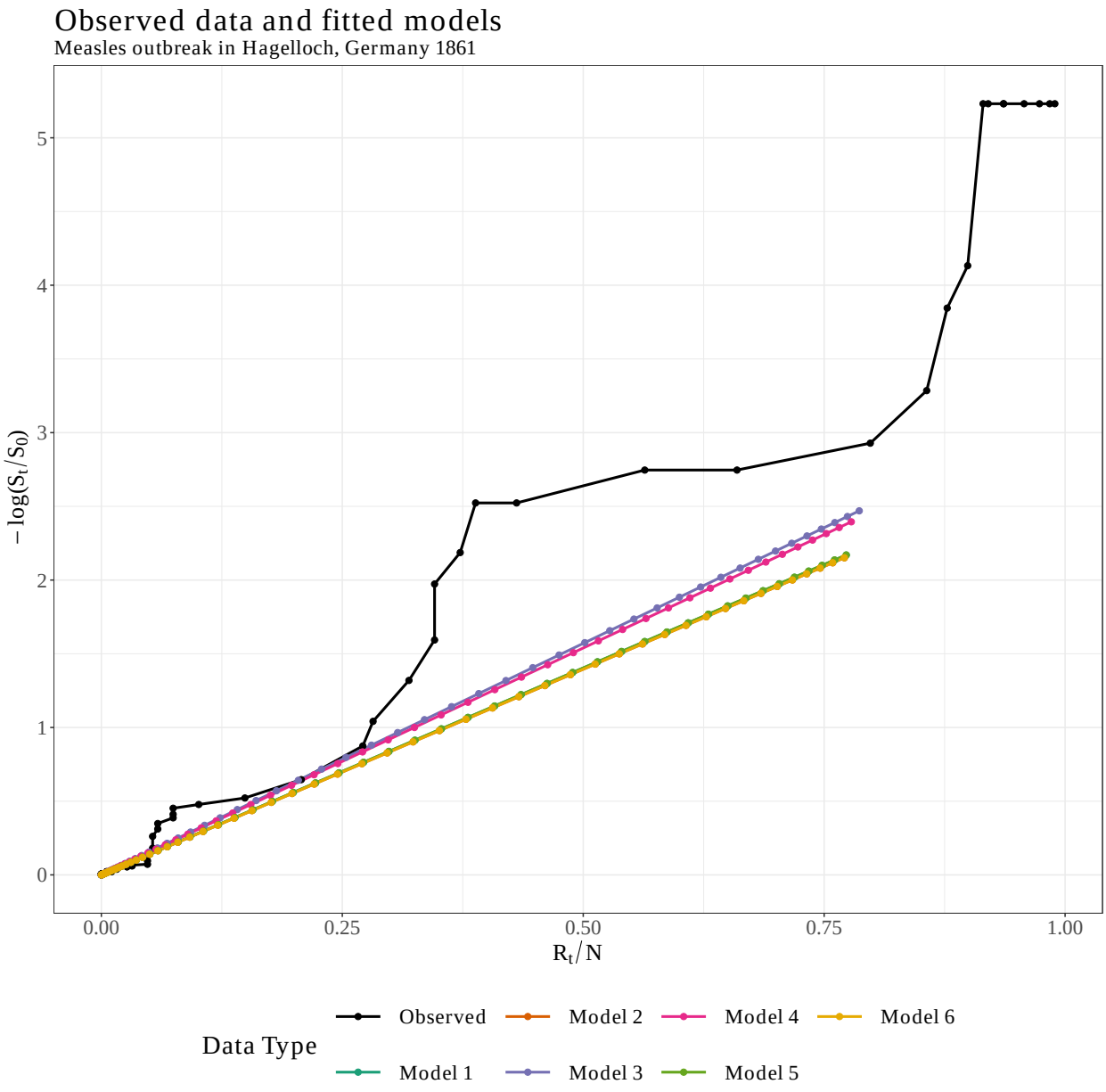

Figure 5.6: Model fits for $K^* = 3$ to the Hagelloch measles data plotted in a log linear transformation.

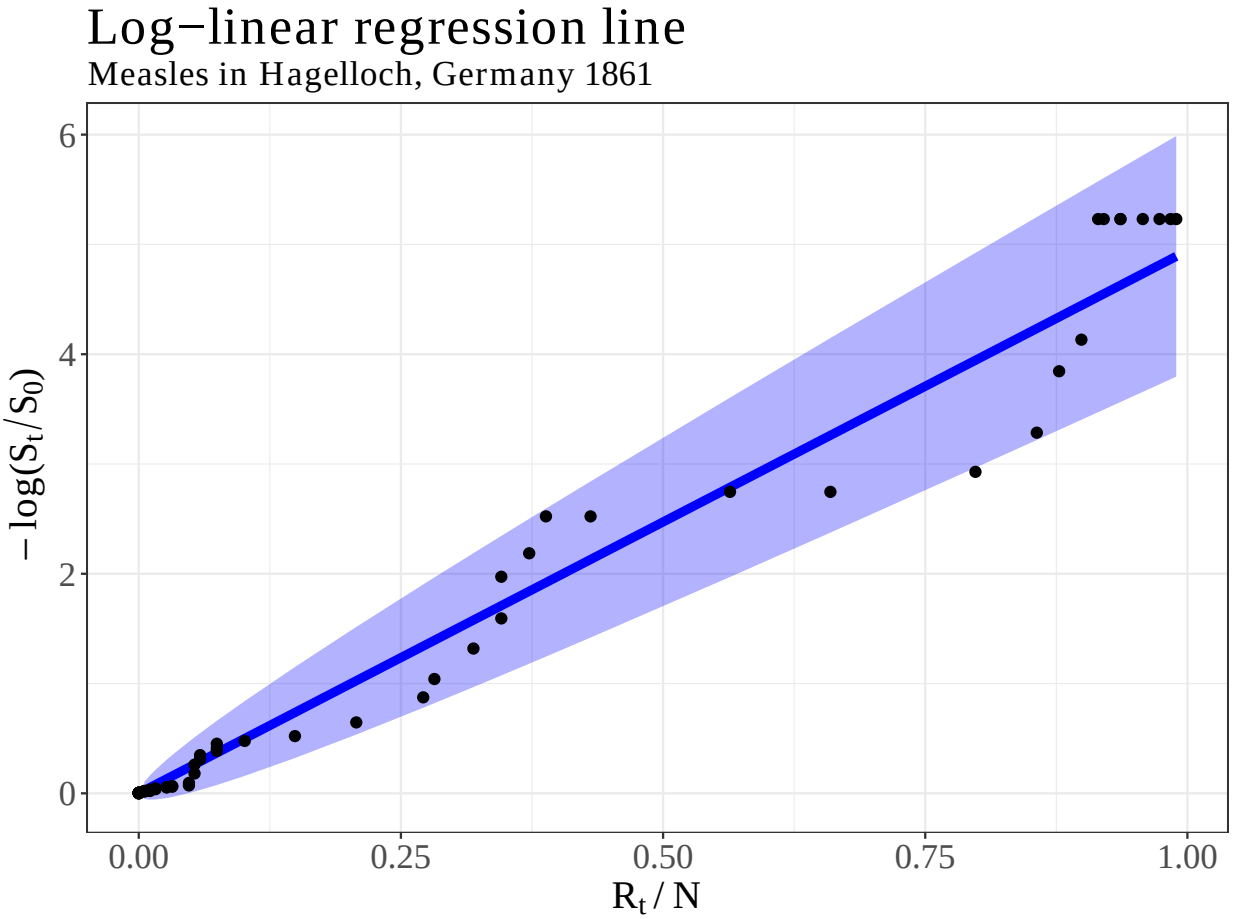

Figure 5.7: Weighted linear regression estimate for observed data with 95% point-wise prediction interval for the observed R_t/N values. The slope of the line corresponds to $\hat{\mathcal{R}}_0$, or roughly, 5.9.

The regular, ternary, and log linear visualizations allow us to see different features of the data and our model estimates over time. The regular view, in Figure 5.4 allows us to show the model fit for each of the three different states where time is the x-axis. However, it is difficult to assess the model fits as a whole as opposed to state-wise.

The ternary view in Figure 5.5 emphasizes the bimodal infectious curve and shows that the SIR model estimates do not fit this bimodal structure well. Conversely, it is more difficult to visualize time in this view, which is important because points may have similar locations in the ternary plot that occur at different time points. With the ternary plot, we see more striking differences between the MSE-fit and the log likelihood-fit models.

Finally, the log-linear estimates in Figure 5.6 are poor fits to the data. Again, it is more difficult to visualize time in this view. Moreover, the weighted linear regression estimate in Figure 5.7 has a much higher estimate of $\hat{\mathcal{R}}_0$ than the other models but seems to fit the data much better.

Overall, the three visualizations show that the $K^* = 3$ SIR model is a poor fit to the Hagelloch data. Therefore, we need to explore different groupings within the population of Hagelloch.

5.2.3 Models for when $K^* > 3$

While in the previous section we determined that $K^* = 3$ was not a sufficient number of states to adequately model the outbreak of measles, in this section we examine models with $K^* > 3$ total states, which are found by partitioning the population into sub-groups. While $K^* = 3$ is the lower bound for the total number of states needed to model the population, the other extreme is $K^* = 3N$ states, which corresponds to an S, I, and R state for every agent in the population. We use the other extreme of $K^* = 3N$ to guide our search for the optimal number of states.

For $K^* = 188 \times 3$, we fit (β_n, γ_n) for the following $n = 1, \ldots, 188$,

$$\begin{aligned} \arg \max_{(\beta_n,\gamma_n)_{n=1,\ldots,N}} \mathcal{L}\left((\beta_n,\gamma_n)_{n=1,\ldots,N};, \mathbf{U}_1, \ldots, \mathbf{U}_N\right) &= P(\mathbf{U}_n = \mathbf{u}_n, \ldots, \mathbf{U}_N = \mathbf{u}_N) \\ &= P(\mathbf{X}) \prod_{n=1}^{N} P(\mathbf{U}_n = \mathbf{u}_n | \mathbf{X}) \end{aligned} \tag{5.5}$$

where $P(\mathbf{U}_n = \mathbf{u}_n)$ is defined in Eq. (5.4). Initially, we use $p_{tn}(\theta)$, a non-random variable, as the probability of infection where

$$p_{tn} = \beta_n \frac{I(t)}{N}, \tag{5.6}$$

where $I(t) = \sum_{n=1}^{N} X_{2n}(t)$ is the total, *expected* number of infectious individuals at time t. We then must jointly estimate (β_n, γ_n) for all n, which is a difficult optimization problem. Instead, we assume $\mathbf{X}$ is known

(i.e. we use the aggregate $\mathbf{u}_n$ as $\mathbf{X} = \mathbf{x}$). This allows us to maximize $P(\mathbf{U}_n = \mathbf{u}_n|\mathbf{X})$ independently from one another. The results of this estimation are shown in Table 5.6. The results are summarized in Table 5.4, Model 7, and the individual estimates of $(\hat{\beta}_n, \hat{\gamma}_n)$ are plotted in Figure 5.8.

The parameter estimates are shown in Figure 5.8. The estimates have mean $(\bar{\beta}_n, \bar{\gamma}_n) = (0.54, 0.13)$ with

$$\hat{\Sigma} = \begin{pmatrix} 0.11 & -0.001 \\ -0.001 & 0.001 \end{pmatrix}.$$

Looking at Figure 5.8, the density of (β_n, γ_n) seems multi-modal and so the mean is not a good summary of the distribution. The total log likelihood is -1165. If we adjust for the number of parameters estimated $(188 \times 2 - 1)$ then the AIC is -2 $\times$ 1540. We use these parameter estimates to guide our search for K^*.

To recap, we estimate individual $(\hat{\beta}_n, \hat{\gamma}_n)$ parameters for each agent and plot the results as β vs. γ in Figure 5.8 along with the 2D density estimate. In this graph, we see two groupings of β and γ estimates. We find that this grouping has a close correspondence to the more natural split of clustering individuals whether they were infected before day $t = 25$, as shown in Figure 5.8. The grouping of being infected before or after day $t = 25$ is displayed on the original household grid in Figure 5.9.

As a result, we focus on models where we use this natural partition of whether agents were infected before or after day $t = 25$. We also find a difference in agents before and after day $t = 15$ so we explore that partition as well. In general, we assume there are G groups of individuals where if an agent belongs to group g then it has parameters (β_g, γ_g). To find estimates of (β_g, γ_g), we maximize the likelihood

$$(\hat{\beta}_g, \hat{\gamma}_g)_{g=1,\ldots G} = \arg \max_{(\beta_g, \gamma_g)_{g=1\ldots,G}} \mathcal{L}((\beta_g, \gamma_g)_{g=1,\ldots,G}; \mathbf{U}_{G_1}, \ldots, \mathbf{U}_{G_G}) \tag{5.7}$$

where $\mathbf{U}_{G_g}$ refers to the set of U statistics for the agents either belonging to group g or interacting with agents in group g. In Equation (5.7), we typically set $\mathbf{U}_{G_g} = \mathbf{U}$, meaning that agents in group g interact homogeneously with the rest of the population. However, we can also set $\mathbf{U}_{G_g}$ so that it corresponds to only a subset of the population. For example, if we subset $\mathbf{U}_{G_g}$ to only the agents belonging to group g, then we only use agents in group g to calculate $p_{tg}(\theta)$,

$$p_{tg}(\theta) = \beta_g \frac{I_{G_g}(t)}{N_{G_g}},$$

where $I_{G_g}(t)$ refers to the number of infectious individuals in G_g at time t and N_{G_g} is the total number of agents in group G_g.

The model with both the maximum log likelihood and minimum AIC is model 11 which has $K^* = 9$ states, $G = 3$ groups and a log likelihood of -1041. The number of parameters estimated in model 11 is 11 (six for (β_g, γ_g), two for split 1 and split 2, and three for U_{G_g}), which brings the AIC/2 to 1052. We see

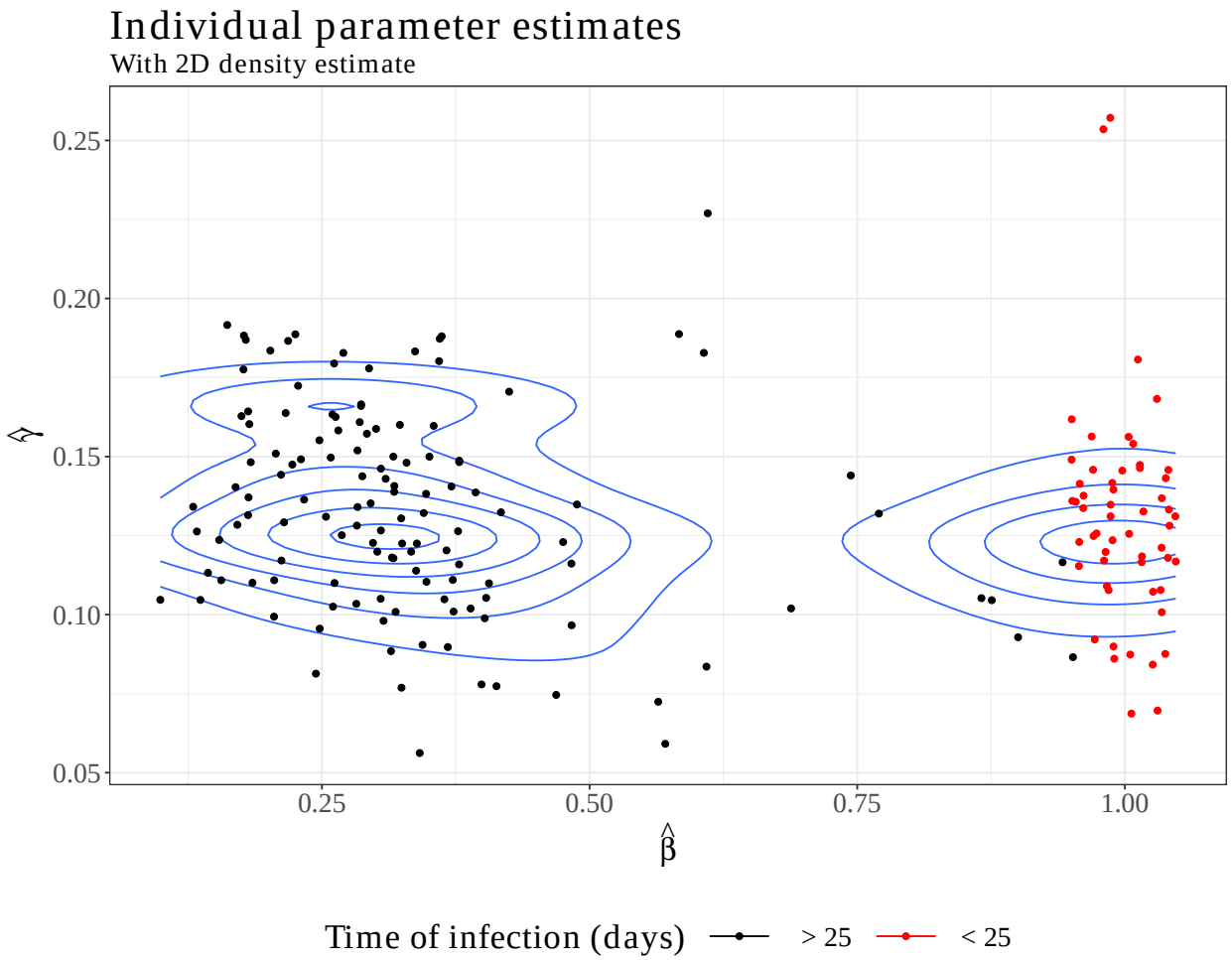

Figure 5.8: Individual estimates of $\hat{\beta}_n$ and $\hat{\gamma}_n$ for each agent $n = 1, \ldots, 188$ obtained by maximizing $\mathcal{L}(\beta_n, \gamma_n; \mathbf{X}, \mathbf{U}_n)$.

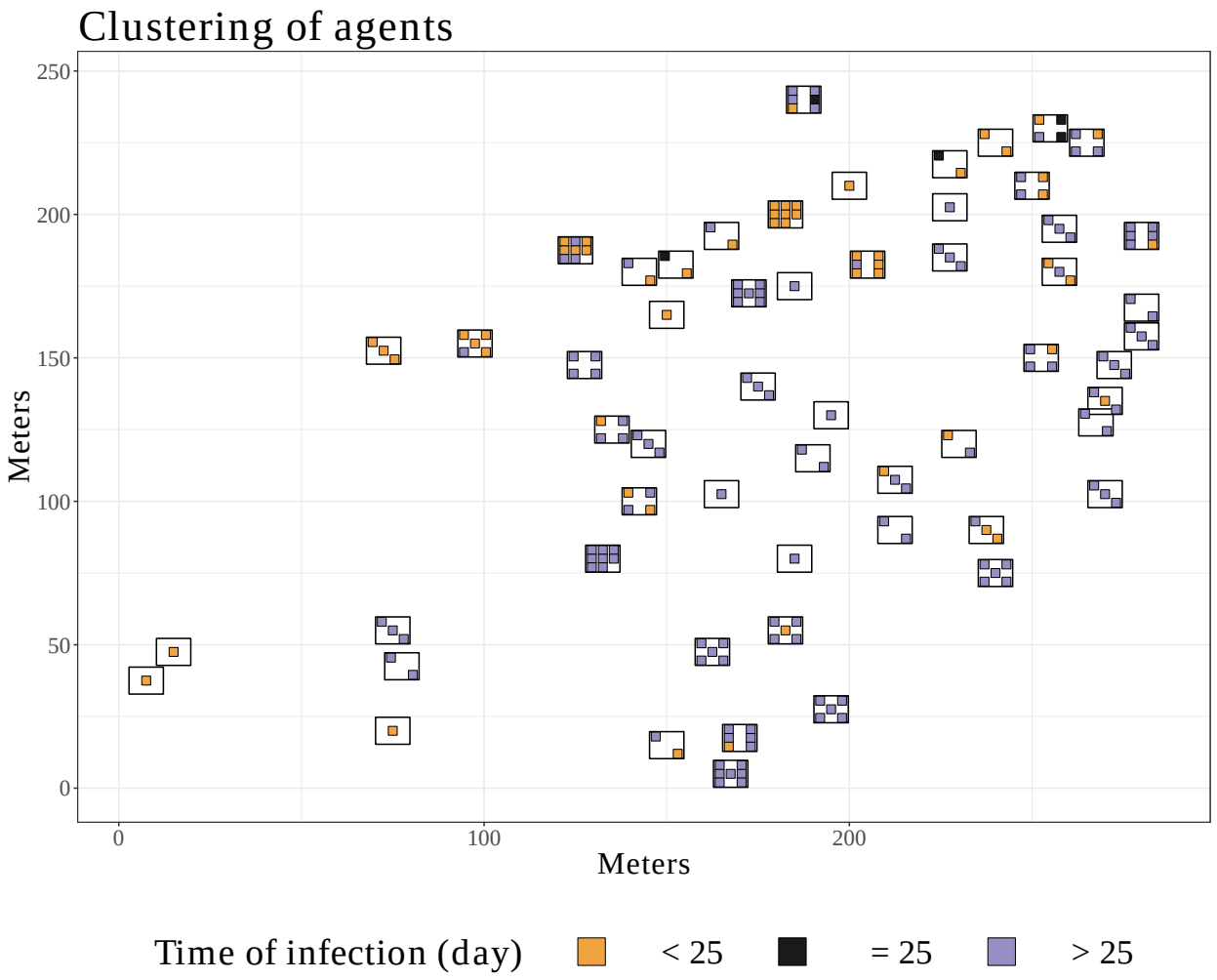

Figure 5.9: Clustering of agents based on time of recorded infection.

Table 5.5: Table of log likelihood and AIC for models with different K^*. Model 1 refers to Model 1 in Table 5.4. Model 7 is the model where all agents have their own (β_n, γ_n). Models 8 and 9 refer to models where $(\beta_n, \gamma_n) = (\beta_1, \gamma_1)$ if the time of infection is before $t = 25$ and $(\beta_n, \gamma_n) = (\beta_2, \gamma_2)$ if the time of infection is after or on $t = 25$. Models 10 and 11 refer to models where $(\beta_n, \gamma_n) = (\beta_1, \gamma_1)$ if time of infection is before $t = 25$, $(\beta_n, \gamma_n) = (\beta_2, \gamma_2)$ if time of infection is in $t = [15, 25)$, and $(\beta_n, \gamma_n) = (\beta_3, \gamma_3)$ if time of infection is in $t \geq 25$.

Model	$\mathbf{U}_{G_g}$	K^*	**Log Like.**	**# Pars.**	**AIC/2**
1	All	3	-1248	3	1251
7	All	564	-1165	396	1561
8	All	6	-1236	7	1243
9	Group g	6	-1103	8	1111
10	All	9	-1229	9	1238
11	Group g	9	-1041	11	1052

Table 5.6: Result of $K^* \geq 3$ SIR model fits to the Hagelloch data.

Model	Random p_t	$\mathbf{U}_{G_g}$	Split 1	Split 2	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\gamma}_1$	$SE(\hat{\gamma}_1)$	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\gamma}_2$	$SE(\hat{\gamma}_2)$	$\hat{\beta}_3$	$SE(\hat{\beta}_3)$	$\hat{\gamma}_3$	$SE(\hat{\gamma}_3)$
1	N	All	NA	NA	0.279	0.008	0.100	0.007	NA	NA	NA	NA	NA	NA	NA	NA
8	Y	All	tI < 25	NA	0.426	0.035	0.102	0.017	0.228	0.014	0.092	0.008	NA	NA	NA	NA
9	N	Group g	tI < 25	NA	0.303	0.016	0.097	0.012	0.569	0.023	0.125	0.010	NA	NA	NA	NA
10	N	All	tI < 15	tI < 25	0.700	0.254	0.310	0.056	0.452	0.043	0.084	0.014	0.244	0.015	0.097	0.009
11	N	Group g	tI < 15	tI < 25	0.348	0.068	0.087	0.034	0.657	0.032	0.111	0.015	0.569	0.023	0.125	0.010

that the two models (models 9 and 11) with $\mathbf{U}_{G_g}$ subset to group g have a much smaller log likelihood than their counterparts with $\mathbf{U}_{G_g}$ as all agents. This is to be expected, because in order to fit the models with completely separate sub-groups, we must set the initial number of infected individuals as known. As such, it is not fair to compare the completely separate group models to the others. However, estimates in these models can be still used to learn about a population such as by conditioning the model on the first, e.g. $t = 25$ days worth of observations.

We plot the estimates in a ternary plot in Figure 5.10. The ternary plot shows that $K^* = 3$ is not a good fit at all, but $K^* = 6$ and $K^* = 9$ are comparable to one another. We also see that only when we assume we have heterogeneous interaction of agents (models 9 and 11), do we see the models better fit the points. We see these two best estimates with a 95% CI (Models 9 and 11) plotted in Figure 5.11. In particular, we see that model 9 fits the data very well.

Since the partitions are intuitive, i.e. it is reasonable to believe the population behaved differently before and after $t = 25$ days, the model log likelihood is relatively large, and especially because the model estimates in Figure 5.10 seems to fit the observed data well, we decide to use the parameter estimates in Models 8 and 9 in our paired stochastic AM.

In particular, for model 8 we estimate the $\mathcal{R}_0$ values for the two groups as 4.17 (95% CI: [3.89, 4.46]) and 2.49 (95% CI: [2.37, 2.62]). For model 9 we estimate the $\mathcal{R}_0$ values for the two groups as 3.13 (95% CI: [2.94, 3.31]) and 4.57 (95% CI: [4.41, 4.72]). This means that depending on whether we consider our

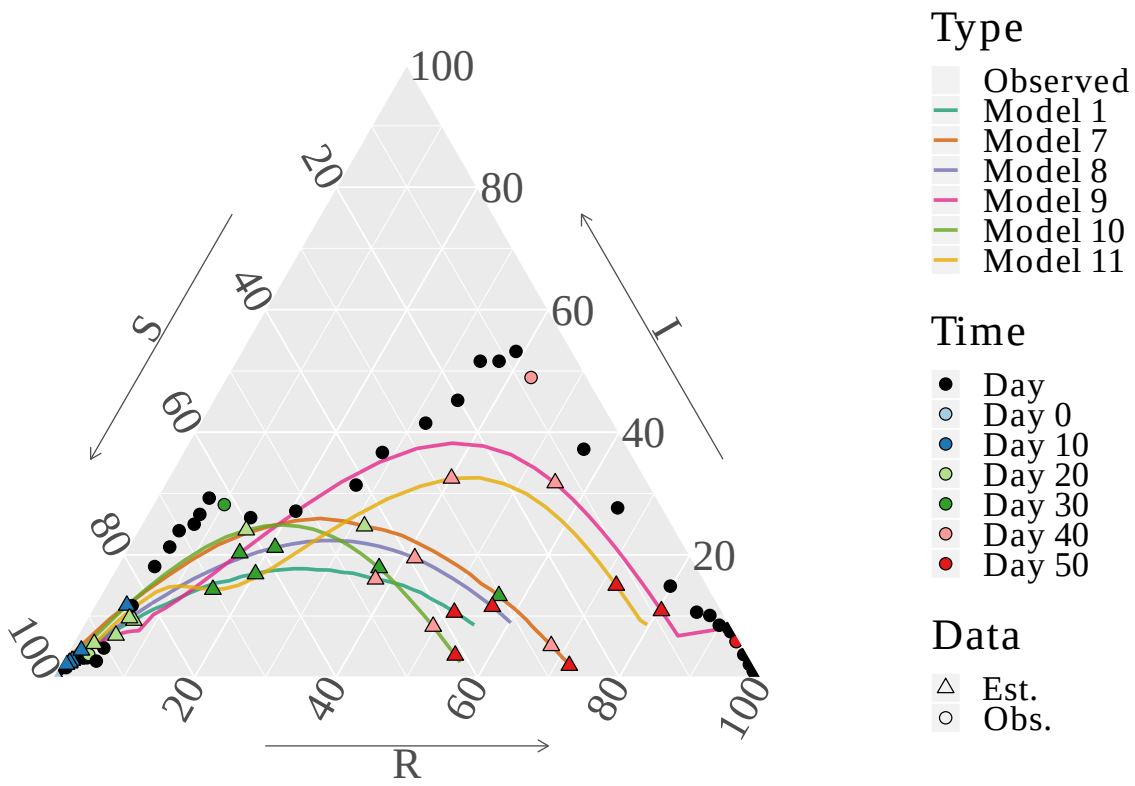

Figure 5.10: Ternary plot of the model estimates in Table 5.5. Every 10th day is filled in with the same color to get a better sense of the time dimension of the epidemic. The observed values are plotted as circles and the estimated values as triangles.

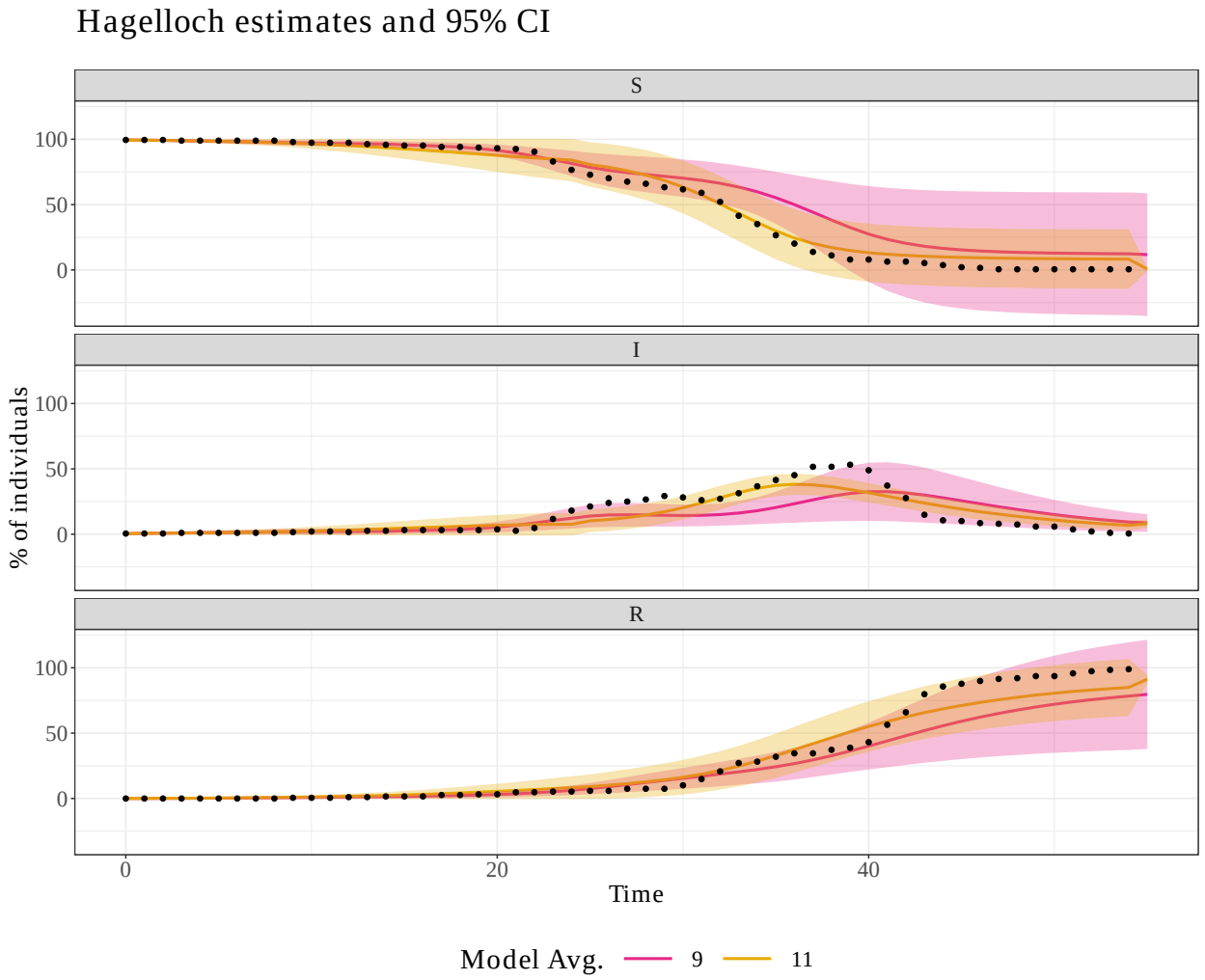

Figure 5.11: Number of agents in each state vs. time, faceted by the aggregate S, I, and R states respectively. The model estimates and their 95% point-wise CIs are shown along with the original observations.

groups of agents to interact homogeneously or not at all has a large role in determining our estimates of how infective measles is. In fact, which group is more infective depends on this assumption.

Moving forward, we will use the estimates from both models 8 and 9 which correspond to the SIR-disease level states with $K^* = 6$ total states. Model 8 assumes of the homogeneous interaction of the sub-groups of agents whereas model 9 assumes that the groups of agents only interact with those in the same group.

One may wonder why we are using time of infection to partition the agents as this was not known *a priori*. In the data, we do know there is an event that occurred around day 25 where purportedly one student infected 26 of his classmates, which we believe is an influential event.

We note that the models we are selecting are possibly not the "best" models, and future work will be dedicated to selecting optimal parameters. However, due to both the "common sense" and statistical reasons mentioned above, we believe that models 8 and 9 adequately capture the necessary population interactions needed to spread the measles through the population. In the next chapter, we will examine hypothetical scenarios where we implement prevention methods into the population.

5.3 Chapter summary

In this chapter, we looked at a data application for our stochastic CM-AM pairs, the Hagelloch measles outbreak of 1861. The outbreak consists of 188 cases over the course of ∼90 days. Using our theory presented in Chapters 2 and 3, we select two SIR-CMs that adequately model the epidemic. In the next two chapters, we use the parameters estimated here in our stochastic CM-AM pair to consider hypothetical issues such as reducing the infectivity of a disease and school closures.

Specifically, in this chapter: we (1) identify different sub-groups of agents, (2) determine the minimal number of states needed to create CM-AM pairs, (3) estimate parameters, and (4) estimate $\mathcal{R}_0$, the reproduction number.

To identify different sub-groups of agents, we conduct EDA on the data set, which is publicly available online via the `R surveillance` package. We look at the agents in terms of their infection and recovery times, spatial characteristics, and demographic characteristics such as which school class the child belongs to. From this, we ascertain that the SIR disease-level states are most appropriate to model the data due to both scientific knowledge of how measles is transferred and practical limitations in the data.

With our disease-level states fixed, we then determine what is our best K^* value, the minimum number of states required. We study a variety of models for $K^* = 3$, which is the case when we assume that the population interacts homogeneously and agents are homogeneous within states. We use two formulations for estimates of probability of infection, study two sufficient statistics that can be used to estimate parameters (distinguishable vs. indistinguishable agents), and two methods of parameter estimation: maximum likelihood and maximizing the sum of square error. We find that (1) these models result in similar

estimates for SIR curves and (2) none of these methods fit the data very well. We assess (1) and (2) using both quantitative and qualitative diagnostics. Especially for issue (2), we introduce ternary and log-linear formulation plots designed specifically to assess the SIR model with $K^* = 3$ states. Using these, we can definitively claim that the $K^* = 3$ is not an adequate number of states to adequately model the outbreak.

We then use the results from $K^* = 3$ models to help us guide our search for the best K^*. In particular, since the $K^* = 3$ models yield very similar results, we confine our remaining models to using log likelihood fits, distinguishable agent sufficient statistics, and the K&M probability of infection formulation. We first look at the other extreme K^*, when $K^* = 188 \times 3$, which corresponds to having a a S, I, and R state for each agent. In this model, we estimate individual infection and recovery parameters, (β_n, γ_n). We then cluster the (β_n, γ_n) plots and find that by splitting the agents at time of infection before and after day $t = 25$ corresponds to the groupings that emerge in the plot well, and note that there may be another partition of the agents at infection time $t = 15$. As such, we explore models where $K^* = 6$ and $K^* = 9$ which corresponds to dividing the population into 2 and 3 groups, respectively.

We explore models with $K^* = 6, 9$ looking at cases where the groups of agents interact homogeneously and where the groups of agents have no interaction outside their sub-groups. We find that both quantitatively and qualitatively, having two groups ($K^* = 6$) of agents outperforms having $K^* = 3$ states and $K^* = 188 \times 3$ states. The models with $K^* = 6$ states also have comparable results to $K^* = 9$ models, despite having a larger AIC. As a consequence, we determine that $K^* = 6$ is an adequate number of states to model the epidemic. In particular, we find that when we consider the sub-groups to be separate from one another, we have a particularly good fit in terms of our visual diagnostics. We will explore both selected models and their CM-AM pairs in the following chapters.

Finally, we assess the value of $\mathcal{R}_0$ for this measles outbreak. As a population estimate (i.e. when we consider all agents to interact homogeneously and have the same rates of infection and recovery), we estimate $\mathcal{R}_0 = 4.94$ (95% CI: [4.68, 5.21]), which is smaller than other $\mathcal{R}_0$ estimates for the measles but is still very large overall compared to other diseases (Gallagher et al., 2019) and is in line with a more recent estimate of $\mathcal{R}_0$ from Getz et al. (2016). When assuming there are two sub-groups of agents that interact homogeneously, we have estimates of $\mathcal{R}_0$ for the two groups of 4.17 (95% CI: [3.89, 4.57]) and 2.49 (95% CI: [2.37, 2.62]), respectively. When we assume the two sub-groups of agents that are considered to have completely separate interactions, we estimate the $\mathcal{R}_0$ values for the two groups as 3.13 (95% CI: [2.84, 3.41])) and 4.35 (95% CI: [4.23, 4.48]). Either way, we see evidence of a difference of infectivity of the measles in the two groups. However, which group is more infectious than the other depends on how we assume the population interacts.

In the next chapter, we will use the above parameter estimates in our CM-AM pair. We will use this CM-AM pair to address questions such as what would have happened if the infectivity of the disease was reduced, what if infected households were isolated, and what if schools were closed.

Chapter 6

Measles: reducing infectiousness

6.1 Introduction

In Chapter 5, we used the observed data in the 1861-1862 Hagelloch measles outbreak to estimate disease parameters and two corresponding (simple) agent-interaction structures. Now, we use these estimated parameters to examine hypothetical scenarios, the results of which policy makers could use to plan responses.

We examine the results of three hypothetical scenarios using our stochastic CM-AM pair:

1. Reduction of the total infectivity disease parameter(s) ($\hat{\beta}_k$)
2. Isolation and quarantine of agents to their households
3. School closure based on a threshold of infectious individuals.

In this chapter, we examine issue (1) in depth, and in the following chapter we examine issues (2) and (3). We split results of the CM-AM in this way because the first issue is more of an abstract concept whereas the second and third issues are concrete prevention measures to be implemented.

There are many ways a modeller could answer the above questions, using many different models, but an accessible and flexible way to examine all three scenarios with one model is through a stochastic CM-AM pair. Since AMs are based on "verbal argumentation," they are easier to understand to the non-computer or data scientist compared to more mathematical models (Epstein, 2007).

As we mentioned, the reduction of the infectivity disease parameter(s) $\hat{\beta}_k$ is more abstract than the other two issues which involve more tangible prevention methods. The idea behind this analysis is to examine possible outcomes of "broad-stroke" prevention measures, regardless of what they may be. Results of this analysis may be presented in the form of statements such as if we can reduce the purported infectivity parameter by $x \pm C_x\%$, we expect to see a $y \pm C_y\%$ reduction in the final size of the outbreak. As such, this analysis of $\hat{\beta}_k$ can be used directly in risk and cost analysis. In the following sections we will

1. Formulate and describe how our CM-AM is changed to simulate prevention strategies
2. Examine homogeneous as well as heterogeneous agent interactions for the models selected when $\hat{\beta}_k$ is reduced at times $t = 0$ and $t = 25$
3. Determine how, generally, the homogeneous model changes when $\hat{\beta}_k$ is reduced at different times.

6.2 Reducing the infectivity parameter $\hat{\beta}_k$

Recall, $\hat{\beta}_k$, is associated with the probability of infection. In the Reed-Frost equations, it is exactly the probability of obtaining an infection from an infectious individual. In the K&M equations, $\beta \times \frac{I(t)}{N}$ is the probability that a susceptible individual will become infectious from one time step to the next. As such, a reduction in β will directly lead to a smaller chance of infection and $\mathcal{R}_0$.

The parameter β may be reduced through a variety of disease prevention and intervention routines: vaccinations, isolation and quarantine, school shut downs, and awareness campaigns (Chen and Jamil, 2006; Grefenstette et al., 2013; Lima et al., 2015; Henao-Restrepo et al., 2017). Agent-based modelers and epidemic modelers, in general, are interested in sensitivity analysis of parameters such as β (Lash and Fink, 2003; Epstein, 2007; Capaldi et al., 2012).

We begin using our stochastic CM-AM pair docked with our estimates $\hat{\beta}_k$, $\hat{\gamma}_k$ obtained from Chapter 5. We then scale $\hat{\beta}_k$ according to the tuning parameter $\rho \in [0, 1]$. This analysis is asking the question: if we reduce our estimate of $\hat{\beta}_k$ to $\rho \cdot \hat{\beta}_k$, then how would the resulting epidemic change? We analyze this question now.

In Chapter 5, we selected two sets of parameters for two different models corresponding to $K^* = 6$ total states: the first set, which assumes that the population interacts homogeneously and the second set, which partitions the population into two sub-populations that have no cross-partition interaction. We will also consider conditioning on the first 25 days of data, as this the natural partition for the two sub-populations. In general, it is possible (and in fact likely) for interventions to be implemented mid-outbreak. We analyze the result of the epidemic for different days when first reduce $\hat{\beta}_k$, which corresponds to interventions being implemented on different days.

To summarize the results of our simulations, we examine (1) peak infectious percent, (2) day of peak infectious percent, (3) final size of the epidemic (i.e. the total number of individuals infected throughout the course of the disease), and (4) infection duration, which are commonly used to measure the severity of diseases (Nishiura and Chowell, 2009; Brooks et al., 2015). The first two summaries represent the "worst" part of the epidemic, the day when the most agents are infectious, and the percent of the population that is infectious on that day, respectively. These worst parts result in heavy burdens on the population including

schools, workplaces, and healthcare facilities. The last two summaries look at how long the epidemic affects the population and how much of the population is infected, respectively.

6.2.1 Analyzing the epidemic from time $t = 0$ onward

We first examine the stochastic CM-AM pair where the agents are separated into two groups with different infection and recovery rates but interact homogeneously with one another. Group 1 is associated with $(\hat{\beta}_1, \hat{\gamma}_1) = (0.43, .10)$ and group 2 is associated with $(\hat{\beta}_2, \hat{\gamma}_2) = (0.23, 0.09)$. We consider the initial states to be known, meaning that one agent in group 1 is infectious at time $t = 0$ and the remaining agents are susceptible. We use our `R` package `catalyst` to simulate the results of the stochastic CM-AM pair. We scale the $\hat{\beta}_k$ values by ρ for $\rho = 0.1, , 0.2, \ldots, 1.0$ and run the AM for each set of scaled parameters for $L = 1000$ simulations.

We plot summaries of the simulations in Figure 6.1. In each of the three figures, the x-axis is the peak % of infectious agents. The (x, y) observations are plotted along with an ellipse where the major and minor axes represent 95% marginal CIs of the variable. In each figure, the observations are colored by the value of ρ, which is used to represent a reduction in $\hat{\beta}_k$. Finally, each figure is grouped by the agent interaction type: heterogeneous or homogeneous.

In the top figure, we plot day of peak % infectious vs. peak % infectious. We first look at the heterogeneous interactions (left). When $\rho < 0.35$, the day of peak infectious is estimated to be within the first month (although this is highly variable for $\rho = 0.2, 0.3$) and the expected peak infectious percentage is less than 10%, which indicates that the outbreak would die off quickly. When $\rho \geq 0.35$, $\hat{\beta}_k$ are large enough to sustain an infection over the course of time shown, on average. One point of interest is that as $\rho > .35$ increases, the sample error for day of peak decreases while the sample error for peak infectious percentage increases. The correlation of the sample errors of day of peak and peak infectious is shown in Table 6.2. The correlation between the sample errors of day of peak infectious percentage and final size is slightly negative. We see that the day of peak % infectious is never below day 25, and that is simply due to the initialization of our model. Since the populations do not interact, we need initial agents from the sub-population that were infectious at time $t = 25$ (as recorded in the data) in order simulate the spread of disease.

For the homogeneous interactions in Figure 6.1 (top right), we see that summary estimates are much more variable compared to the heterogeneous interaction models, which we would expect. We see many of the same trends as in the heterogeneous interaction figure. The correlation between the sample errors is -0.09, shown in Table 6.1.

Overall, in the top images, we see that reduction in $\hat{\beta}_k$, on average, leads to a reduction in peak % of infectious agents. For day of peak % infectious, an increase in ρ, on average, leads to an increase of day

Table 6.1: Correlation of variance of summary variables from AM simulations for homogeneous model for t=0 onward.

	Final Size	Peak Infectious %	Day of Peak Infectious %	Infection Duration
Final Size	1.00	0.95	0.06	0.74
Peak Infectious %	0.95	1.00	-0.09	0.57
Day of Peak Infectious %	0.06	-0.09	1.00	0.71
Infection Duration	0.74	0.57	0.71	1.00

Table 6.2: Correlation of variance of summary variables from AM simulations for the heterogeneous model for t=0 onward.

	Final Size	Peak Infectious %	Day of Peak Infectious %	Infection duration
Final Size	1.00	0.99	-0.63	-0.31
Peak Infectious %	0.99	1.00	-0.64	-0.32
Day of Peak Infectious %	-0.63	-0.64	1.00	0.19
Infection duration	-0.31	-0.32	0.19	1.00

of peak % infectious up to $\rho = 0.35$. Values of $\rho > 0.35$, on average, lead to a decrease of day of peak % infectious. This implies that a reduction in $\hat{\beta}_k$ can actually result in prolonging the outbreak.

In Figure 6.1 (middle) we plot final size (% of total infected over course of epidemic) and peak infectious percentage for the different values of ρ. For both the heterogeneous and homogeneous agent interactions for the given ρ, there appears to be, on average, logistic growth of final size given the peak % infectious. Again the homogeneous interaction model is much more variable than the heterogeneous interaction model. The correlation of the sample error between these two variables is 0.95, which means that large uncertainty for infection duration is very strongly associated with large uncertainty for peak % of infectious and *vice versa*. This indicates that it is very difficult to predict the "worst" day of the measles epidemic, regardless of ρ, especially assuming homogeneous agent interaction.

We can conclude that if we want to reduce the final size of the epidemic to 75% of the population, we need to reduce $\hat{\beta}_k^* < 0.5 \cdot \hat{\beta}_k$, which is a very sizable reduction! However, if we can decrease $\hat{\beta}_k^* < 0.4 \cdot \hat{\beta}_k$, we can reduce the final size to 50% of the epidemic and if $\hat{\beta}_k^* < 0.3 \cdot \hat{\beta}_k$, then the final size is going to be less than 25% of the total population with 95% certainty.

In Figure 6.1 (bottom) we plot infection duration (number of days until there are no more infectious individuals) and peak infectious percentage for the different values of ρ. Once again, we see the similar trends for the heterogeneous interaction and homogeneous interaction models where the homogeneous interaction model has much larger CIs than the heterogeneous interaction model. We see that the patterns mirror those in Figure 6.1 (top). Again, for the heterogeneous interaction model, infection duration is at least 25 days due to the initialization of the model.

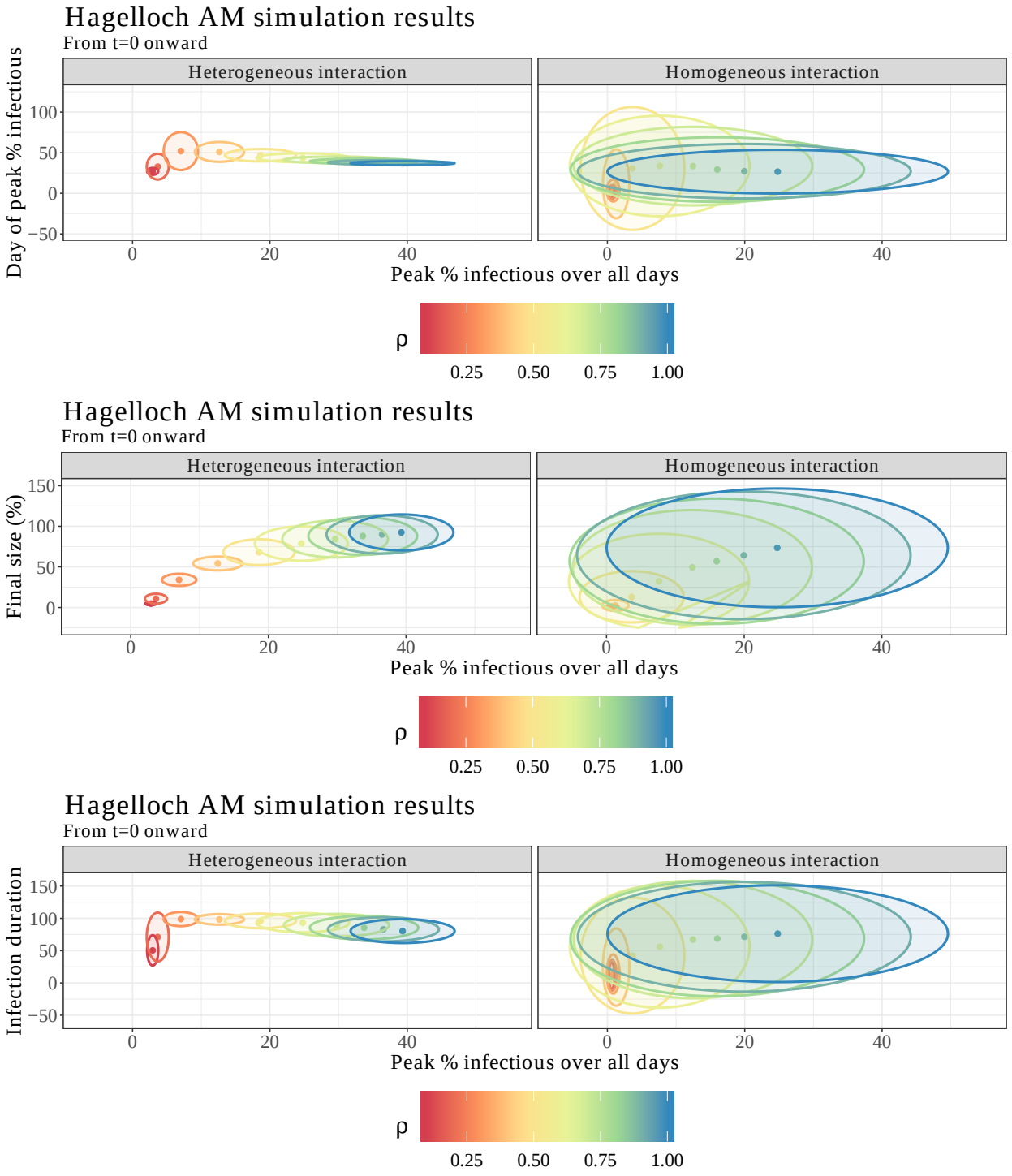

Figure 6.1: Top: day of peak infection and peak infectious. Bottom: final size and peak infectious. Results of AM simulation and 95% CIs. One AM consists of two groups of agents who interact across the groups (homogeneous) and the other does not interact across groups (heterogeneous). Each AM was run 1000 times with $\hat{\beta}_1 = \rho \times 0.43, \hat{\beta}_2 = \rho \times 0.23, \hat{\gamma}_1 = \rho \times 0.10, \hat{\gamma}_2 = \rho \times 0.09$.

From these results, we can conclude that if we can reduce $\hat{\beta}_k$ to 50% of its estimated value, then we can reduce the expected final size of the epidemic to 50% of the population. If we were to reduce it to 70% of its estimated value, we would only expect to see a final size reduction to about 80% of the population. If we reduce $\hat{\beta}_k$ to between 30-40% of its initial value then we would actually expect to prolong the epidemic in the sense that we would continue to see measles cases for a longer period of time than if we did nothing to mitigate the epidemic.

The primary conclusion is that from the start of the epidemic ($t = 0$), we would have needed a drastic reduction in $\hat{\beta}_k$ if we were to have any chance of stopping a near-population wide epidemic. We will next examine what reductions in $\hat{\beta}_k$ are required when preventions are not implemented until time $t = 25$.

6.2.2 Analyzing the epidemic from time $t = 25$ onward

Another situation to explore is when prevention policies are implemented part way through an epidemic. In particular, we examine the population 25 days after the initial case was recorded.

The time $t = 25$ is chosen because it corresponds nicely to the partition of the two sub-groups of agents found in Chapter 5. Admittedly, this partition would not be known *a priori* since we are not grouping the population by a demographic characteristic and future work will be dedicated to selecting a t^* in which we assume the previous data is known.

Still, we can examine how an epidemic changes by varying values of $\hat{\beta}_k$. We use the same estimated parameters as in Section 6.2.1. The difference is that we set the *initial number infectious* in each group (40 and 5, respectively) according to their state on day 26. The results are shown in Figure 6.2.

The results are similar to attempting to reduce $\hat{\beta}_k$ from time $t = 0$. Again, we see that we would like to reduce $\hat{\beta}_k$ by about 50% or more to have significant results on reducing the effect of the outbreak. We see that preventions implemented at $t = 25$ are less effective than having implemented preventions from the beginning but are still useful overall. We also see that we have little variance in the estimate of the day of peak infectious, even dependent on ρ. However, the *peak % infectious* remains a highly variable estimate.

On the other hand, for both the heterogeneous and especially the homogeneous interaction models for $t = 25$ onward, we see that the estimates have much less variation than for $t = 0$ onward, which is to be expected as we are conditioning on the observations up time $t = 25$. The homogeneous model's reduction in variability can also be explained by the significant nature of $t = 25$, which corresponds to how the population is partitioned into two sub-groups. To see a sizable reduction in the final size of the epidemic (over 25%), we would like $\rho < 0.4$. However, once ρ is less than this threshold, the results are seemingly exponential in how much we can stop the epidemic from spreading further!

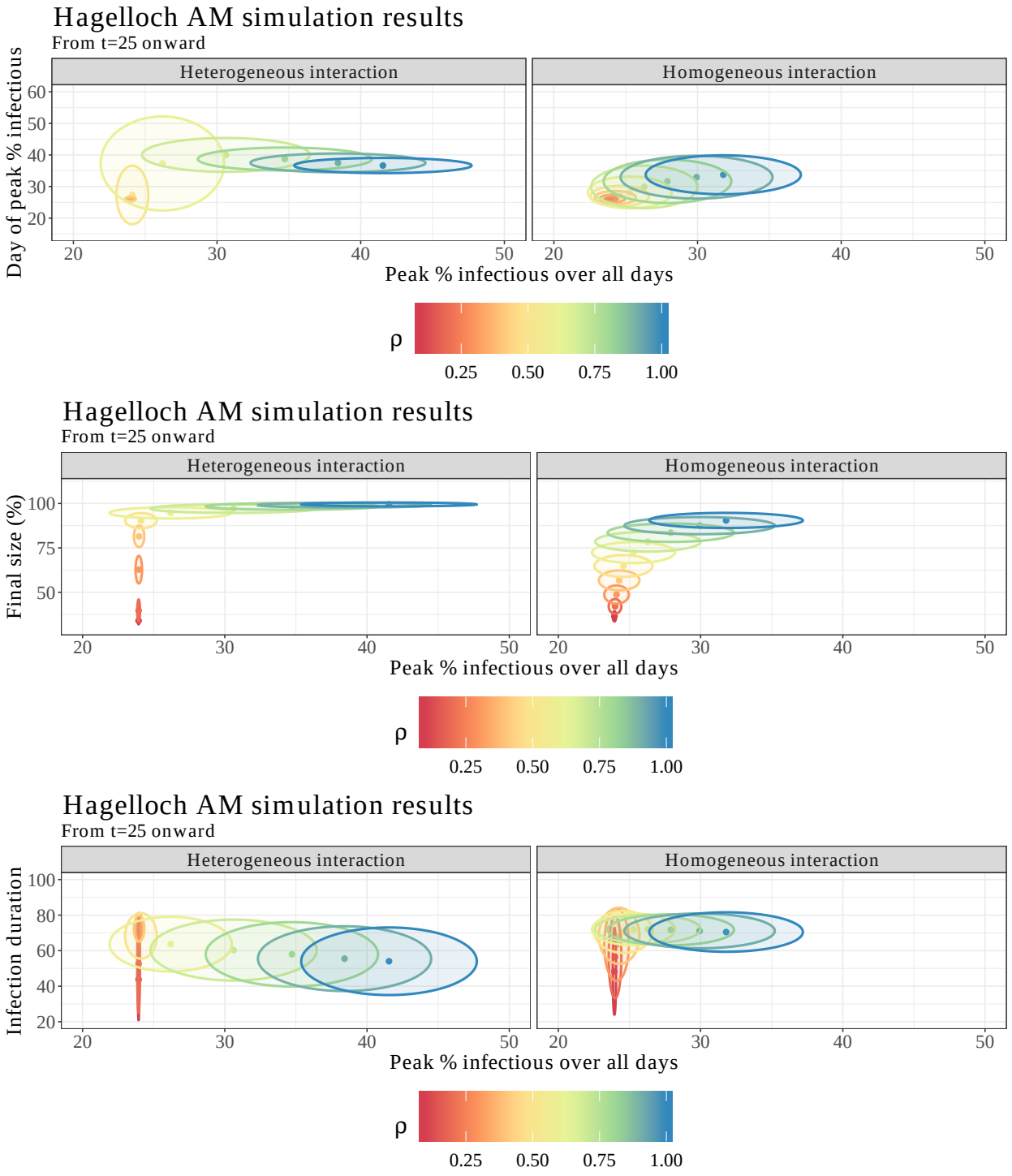

Figure 6.2: Top: day of peak infection and peak infectious. Bottom: final size and peak infectious. Results of AM simulation and 95% CIs. One AM consists of two groups of agents who interact across the groups (homogeneous) and the other does not interact across groups (heterogeneous). Each AM was run 1000 times with $\hat{\beta}_1 = \rho \times 0.43, \hat{\beta}_2 = \rho \times 0.23, \hat{\gamma}_1 = \rho \times 0.10, \hat{\gamma}_2 = \rho \times 0.09$.

6.2.3 Prevention over time

We more generally examine how the final size of the epidemic changes with regards to the day the infectivity parameter, $\hat{\beta}_k$ is reduced. In this analysis, we condition on the first $t-1$ days and begin our simulations on day t. We plot the results in Figure 6.3 which shows *final size* vs. *peak % of infectious* and is grouped by the day the simulations began. We examine ρ values between 0.1 and 0.9.

As the day of reduction increases we find that the the sample errors for *final size* and *peak % of infectious* decrease. This is likely due both to the decreasing size of the susceptible population and the fewer number of days in the simulation. We do note that there is a significant difference between the widths of the 95% CIs from day 6 to day 11, compared to the difference in other days. We see for small values of ρ, the *final size*, on average, increases as the day of reduction increases, which we expect since we are conditioning on the data. There is a large difference in expected final size for small values of ρ between reduction day 21 and 26. In the data, over 20 individuals were infected in that time frame. On day of reduction 36, we see that the outbreak cannot be stopped. However, it seems that up to one month, a reduction in $\hat{\beta}_k$ will, on average, reduce the final size of the epidemic. Again, the reduction of the final size is highly dependent on ρ, with larger values of ρ leading to even greater reductions in final size.

In the case of Hagelloch outbreak, these simulations show that the quicker the response is to the epidemic, the better. Nonetheless, preventions are still worthwhile in terms of reducing the final size of the epidemic even a month after the start, when almost 40% of the population is already infected.

6.2.4 Reducing the infectivity parameter: summary

We examine the hypothetical scenario of what would happen in Hagelloch if we were able to reduce the infectivity of measles, specifically by reducing β. We use the estimates for $\hat{\beta}_k$ obtained in Chapter 5 and then scale them by tuning parameter ρ. To analyze our resulting CM-AM pair simulations, we look at peak % infectious, day of peak % infectious, final size of the epidemic, and infection duration, which are common measures in epidemic theory. We examine these statistics using both heterogeneous and homogeneous interaction of agents. We also examine the results of the epidemic when we reduce the infectivity of the disease at different times t, conditioning on the first $t-1$ observations.

Our results show that, we can better reduce the severity of an epidemic with earlier preventions. Moreover, we find that we need to reduce $\hat{\beta}_k$ by at least 50% to obtain significant reductions in final size of the epidemic. As a result of this analysis, policy makers and scientists would have a goal for which how effective prevention measures should be.

Finally, we examine when we should implement these preventions. Of course, it is better to respond to the outbreak as quickly as possible but Figure 6.3 shows that even a month after the outbreak, we can still reduce the final size the epidemic substantially.

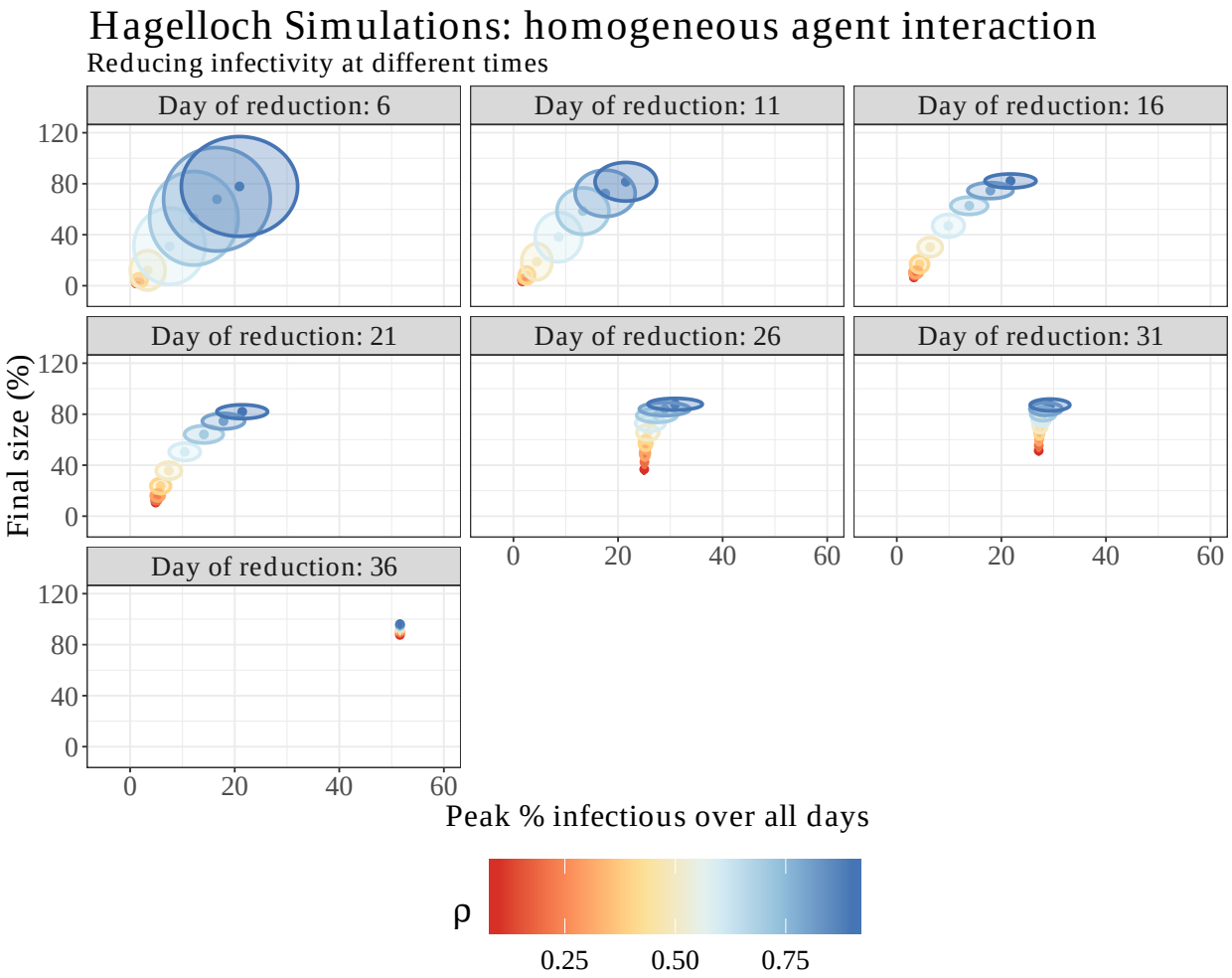

Figure 6.3: Hagelloch simulations with homogeneous agent interaction where we condition on the first $t-1$ data points and include a reduced infectivity parameter, $\rho\hat{\beta}_k$ on day t. Each AM was run 1000 times.

Overall, this analysis of $\hat{\beta}_k$ allows us to quantify how the outbreak might proceed if we were able to implement preventions at a high level. In turn, this can be used to inform policy decisions and cost analysis to determine which preventions may be more practical and affordable for a given population.

In the next chapter we will examine prevention scenarios of a more concrete nature.

Chapter 7

Measles: agent interaction restriction

7.1 Introduction

In Chapter 5, we introduced the Hagelloch measles data set and estimated disease parameters to use in our stochastic CM-AM pair, and in Chapter 6, we analyzed hypothetical scenarios where we looked at the resulting epidemic under the reduction of the infectivity parameter, without specifying how the reduction is accomplished. In this section, we examine more tangible intervention strategies through agent interaction restriction: (i) quarantine and isolation of agents and (ii) school closure.

The above interventions are all commonly used in public health to curb the spread of disease and so should be examined with our CM-AM pair. There are two primary differences between these two tangible intervention strategies and the general reduction of the infectivity parameter in Chapter 6: (1) whereas reduction in the infectivity parameter effected all agents homogeneously, these agent interaction restrictions are local changes, and (2) to more realistically mimic real world epidemics, we use the *observed* number of infectious contacts an agent has instead of their *expected* number of infectious contacts to determine whether an agent becomes infectious from one time step to the next. Statistically, the second issue means that in the Multinomial random variable transition model shown in Eq. (2.6), the probability of becoming infectious is now a random variable. Both of these issues serve to make our models more complex and are especially apparent in producing CIs for our summary estimates.

In this chapter we,

1. Formulate and describe our new stochastic CM-AM pairs with preventions including (i) isolation and quarantine and (ii) school closure
2. Examine the results of isolation and quarantine
3. Determine how the epidemic would have changed under school closure

4. Discuss the advantages and shortcomings of the results in these hypothetical scenarios.

7.2 Formulation and description of CM-AM with preventions

In the stochastic CM setting, recall that each susceptible agent has the following probability of becoming infectious under the K&M framework,

$$A_{t,n}|(A_{t-1,n}=0) = W_{t-1,n,S} \sim \text{Bernoulli}\left(\beta_n \frac{I(t-1)}{N}\right)$$

We emphasize that the probability of transition is non-random as $I(t-1)$ is the *expected* number of infectious individuals under the true model. However, when moving to the AM framework, practically, it makes more sense to have

$$A_{t,n}|(A_{t-1,n}=0) = W_{t-1,n,S} \sim \text{Bernoulli}\left(\beta_n \frac{\hat{\eta}\,(t-1,n)}{N}\right),$$

where $\hat{\eta}\,(t-1,n)$ is the number of infectious contacts agent n has between time $t-1$ and t. When each agent has an equal chance of contacting another agent, then $\hat{\eta}\,(t-1,n) = \hat{I}(t-1)$, the *observed* number of infectious at time $t-1$. As such, the below analysis for both isolation and quarantine and school closure will be inherently more variable than when using a non-random probability of transition. This disparity is worth looking further into and will be addressed in future work.

The first models we introduce are for the isolation and quarantine of agents. According to the US Health and Human Services (HHS 2019), *isolation* is defined as restricting the interaction of *ill* individuals while *quarantine* is defined as restricting the interaction of *well* individuals. More specifically, we isolate an infectious agent after some delay period d to her household where she can only contact other housemates. For quarantine, we keep the entire household of an infectious agent at home after delay period d. The models we analyze are of the form,

$$\begin{aligned}
W_{t-1,n,S} &\sim \text{Bernoulli}\left(\hat{\beta}_n \frac{\hat{\eta}\,(t-1,n)}{N}\right) \\
W_{t-1,n,R} &\sim \text{Bernoulli}\left(\hat{\gamma}_n\right) \\
A_{t,n}|\mathbf{A}_{t-1} &= \begin{cases} 1 + W_{t-1,n,S} & \text{if } A_{t-1,n}=1 \\ 2 + W_{t-1,n,R} & \text{if } A_{t-1,n}=2 \\ 3 & \text{if } A_{t-1,n}=3 \end{cases}
\end{aligned} \tag{7.1}$$

where $(\hat{\beta}_n, \hat{\gamma}_n)$ were determined in Chapter 5. Here, $\hat{\eta}(t,n)$ gives the observed number of infectious contacts of agent n at time t. Isolation of the agent is implemented if $t > t^*_{1,n} + d$ where $t^*_{1,n}$ is maximum time where

agent n is susceptible and d is some period of delay. Let $h(n)$ be the indices of the housemates of agent n. Let $i(t,d)$ be the indices of the agents in isolation at time t after delay period d. Let $q(t,d)$ be the indices of the agents in quarantine at time t after delay period d. Let $j(t)$ be the indices of the infectious individuals at time t. Let $r(t,n)$ be the indices that are to be removed from the contacts of agent n at time t. This can be expressed in set notation as

$$\begin{aligned} h(n) &= \{m \neq n : \text{agent m is housemate of agent n}\} \\ i(t,d) &= \{m : t > t^*_{1,m} + d\} \\ q(t,d) &= \{m : h(m) \cap i(t,d) \neq \emptyset\} \\ j(t) &= \{n : \mathbf{A}_{t,n} = 1\} \\ r(t,n) &= \text{indices of removed contacts for agent } n \text{ at time } t. \end{aligned}$$

Then the number of infectious contacts each agent n has at time t, $\hat{\eta}(t,n)$, may be expressed as

$$\hat{\eta}\,(t,n) = \begin{cases} \#\left((j(t) \setminus r(t,n)) \cup h(n)\right) & \text{if } n \notin r(t,d) \\ \#\, h(n) & \text{otherwise} \end{cases} . \tag{7.2}$$

Equation (7.2) restricts the number of infectious contacts of the infected individual n but always includes the housemates. We examine two sets of $r(t,n)$: one for the isolation routine and one for the quarantine routine,

$$\begin{aligned} r_i(t,n) &= i(t,d) \\ r_q(t,n) &= q(t,d). \end{aligned}$$

In summary for isolation and quarantine routines, $\hat{\eta}(t,n)$ gives the number of infectious neighbors of agent n at time t and accounts for isolation and quarantine routines based on the specification of $r(t,n)$, the indices of agents which we remove from the contact set of agent n at time t. The function $\hat{\eta}(t,n)$ allows us to more naturally model the interaction of agents as we actually simulate the contact and spread of disease based on the agents observed states, as opposed to their expected state.

By adjusting the delay period d, we can analyze various scenarios depending on when the infected agent is isolated or quarantined. When $d = 0$, we assume the agent is isolated as soon as she is infectious. For measles, we expect the infectious period to be between 4 days before and after the occurrence of the measles rash. Therefore, we analyze values of d between 0 and 8. Following the analysis of the isolation routine, we examine quarantine of housemates of infectious individuals to their house, again with a delay d. We discuss the results of this analysis, which are displayed in Figure 7.1.

We also restrict the contacts of agents when implementing school closure. In the Hagelloch data set, there are three classes: pre-school, first class (primary school), and second class (secondary school). Here, pre-school means that a child does not attend school.

To estimate the disease parameters in Chapter 5, we assumed that the agents interacted homogeneously either with the entire population or a sub-group of agents. In this scenario, we remove classmates from the contact list of agents once the total class infection reaches a certain threshold (ρ_1, ρ_2). That said, agents will be free to interact with non-classmates. Arguably, this scenario does not line up with reality, as we may expect children to be more susceptible to their classmates than others. We emphasize, then, that this scenario is conditioned on the assumption that the parameters and interactions inferred from Chapter 5 are true, the evidence being that the parameters provide a good fit to the actual data. In the future, we would like to explore stochastic CM-AM pairs with more complex agent interactions.

Let $c(n)$ be the indices of the classmates of agent n,

$$c(n) = \{\text{m} \neq \text{n: agent m is a classmate of agent n}\}. \tag{7.3}$$

Let ρ_c be the threshold of how many children in class k must be infected simultaneously before we shut down the school. Let T_c be the duration of any school closure. Let $\hat{I}_{t,c}$ be the observed number of infectious children at time t in class c, and let N_c be the number of children in class c for $c = 1$ and 2. We close down school c for the next T_c days if $\frac{\hat{I}_{t,c}}{N_c} > \rho_c$. That is, we assume that classmates no longer contact one another for the duration of T_c days. Let $t_c^* = \min\{\min\{t : \frac{\hat{I}_{t,c}}{N_c} > \rho_c\}, T\}$ be the first day the threshold is crossed (or T if it is never crossed). More formally, the number of infectious contacts at each time step for agent n is given by

$$\hat{\eta}\left(t, n\right) = \begin{cases} \#\left(j(t) \setminus c(n)\right) & \text{if } t_c^* \leq t \leq t_c^* + T_c \\ \# \, j(t) & \text{otherwise} \end{cases} . \tag{7.4}$$

In words, if agent n's school is currently closed down, the number of infectious contacts of agent n is equal to the total number of infectious agents minus the number of classmates agent n has. If agent n's school is not closed down, the number of infectious agents is equal to the total number of infectious agents at time t.

In general, intervention strategies can be implemented by defining $\hat{\eta}(t, n)$, the observed number of infectious contacts at each state. We adjust the number of infectious contacts at time t of agent n,

$$\hat{\eta}(t, n) = \#\left(j(t) \setminus \left(r(t, n) \cup \{n\}\right) \cup p(n)\right) \text{ if } n \in B_b \tag{7.5}$$

where $j(t)$ is the set of indices infectious agents at time t, $r(t,n)$ is the set of indices of agents that are not going to contact agent n at time t (the removal set), $p(n)$ is the set of indices of the permanent contacts of agent n, and B_b is some particular agent state for some space $B = B_1 \cup B_2 \cup \ldots$ and $B_i \cap B_j = \emptyset$ for all i, j.

Eq. (7.5) is a general way of writing the number of infectious contacts for each agent at a given time. In words, we take the set of infectious individuals, take out contacts who are limited by the prevention implemented by the agent being in state B_b but retain permanent contacts such as housemates.

7.3 Isolation and quarantine results

While in Section 6.2 we examined the effects of reducing $\hat{\beta}_k$, a global estimate, we now examine the effects of local interventions such as isolation and quarantine of agents. That is, using our estimates of the disease parameters, we use the stochastic CM-AM pair to analyze local interactions while implementing isolation and quarantine routines with some delay.

For the Hagelloch data set, we implement both isolation and quarantine routines. Specifically, in the isolation routine we note the date of each child's initial infection. After some delay d, we isolate the child to her home for the remainder of her infectious period. Once the isolation is in effect, the child can only spread the disease to the children who belong to the same household. In the extreme case when $d = 0$, we would expect there to be a very slim chance of an outbreak since the child is immediately isolated. A more likely scenario is that the child is not isolated until some $d > 0$ period of time has passed. We investigate a delay period of $d \in \{0, 2, 4, 6, 8\}$ days.

The quarantine routine is similar in that we isolate an infectious child after some delay d. However, we now quarantine some of the child's contacts. In our quarantine routine, we quarantine the children who belong to the same household as the infectious child after delay d. In theory, this should prevent outbreaks that could occur when another child in the household is infectious but not yet isolated, allowing her to spread the disease.

We first compare the general results of isolation of an infectious child after delay period d compared to quarantine of housemates of an infectious child after delay period d (top row of Figure 7.1). As the quarantine routine is a superset of the isolation routine, it is no surprise that the average *final size* and *peak % infectious* are all less than their isolation routine counterparts. However, we find that the *day of peak % infectious* is larger for quarantine routines than their isolation counterparts with the same delay. We also see that the sample error for *peak % infectious*, *final size*, and *peak % infectious* also decrease, and the sample error reduction seems to be more prominent for smaller delay values. This allows us to conclude that quarantine of housemates of the infectious child along with isolation of the infectious child is more effective than isolation alone. For example, for a delay period 0, we expect the difference between isolation

and quarantine final size to be 0.62% less, a delay period of 4 days we expect the difference to be 10% less, and a delay period of 8 days we expect the difference to be 4.52% less.

We can also compare the difference in delay periods to one another. We focus on the results of the isolation routine, but the quarantine routine results are similar. In general we see that a smaller delay, on average, is associated with a smaller average *day of peak % infectious*, *final size*, and *peak % infectious* as well as with a smaller sample error of those same attributes. For a random probability of infection (i.e. the model in Eq. (7.2)) with no isolation, we would expect the final size to be 64% (95% CI: [0, 78]%) of the population; when we have a delay of 8 days, we expect the final size to be 37% (95% CI: [0, 55]%); when we have a delay of 4 days, we expect the final size to be 11% (95% CI: [0, 27]%); and when we have a delay of 2 days, we expect the final size to be 3% (95% CI: [0, 11]%). Therefore, we see that isolating children as soon as possible is much more effective, in the sense that even a 4 day delay can result in an expected final size from 64% to 11%!

Overall, we see that quarantine and isolation can be very effective at reducing the spread of measles on this population of Hagelloch children. Again, we emphasize that this analysis is conditioned on our parameter estimates for $\hat{\beta}_k$, $\hat{\gamma}_k$, and the homogeneous interaction of agents.

7.4 School closure

We analyze the potential effect of shutting down the first and second classes would have on the spread of the epidemic. We examine AM scenarios with $\rho_1 = \rho_2 \in \{0, 0.2, \ldots, 1\}$ with $T_c = T$, which is a permanent school closure once the threshold is met. Astute readers will note that in Eq. (7.4), it is possible to exclude household members from $\hat{\eta}$ if the two belong to the same class. The results, shown below in Figure 7.2, however, show that this choice does not matter much.

Figure 7.2 shows day of peak % infectious vs. peak % infectious (top) and % final size (%) vs. peak % infectious (bottom) where in both graphs, the observations are colored by the closure threshold. The length of the horizontal and vertical axes of the ellipses represent the marginal 95% CIs of the estimates for 1000 simulations for each threshold value.

The results from Figure 7.2 are inconclusive, especially since the ellipses almost all overlap one another. These estimates are highly variable and final size CIs, for example, span the entire range of 0 to 100% for every threshold value. The average estimate for the final size for when $\rho_c = 0$, meaning a permanent school closure as soon as one child in the class is infectious, seems to be smaller than the other threshold values, but again this is not a significant difference.

One thing we see is that the average estimate of final size is larger for larger closure thresholds up until $\rho_c = 0.4$. Then, we see that a closure threshold of $\rho_c = 1$ has a smaller final size estimate than for both $\rho_c = 0.6, 0.8$. The threshold $\rho_c = 1$ corresponds to never shutting down the school and should (and does)

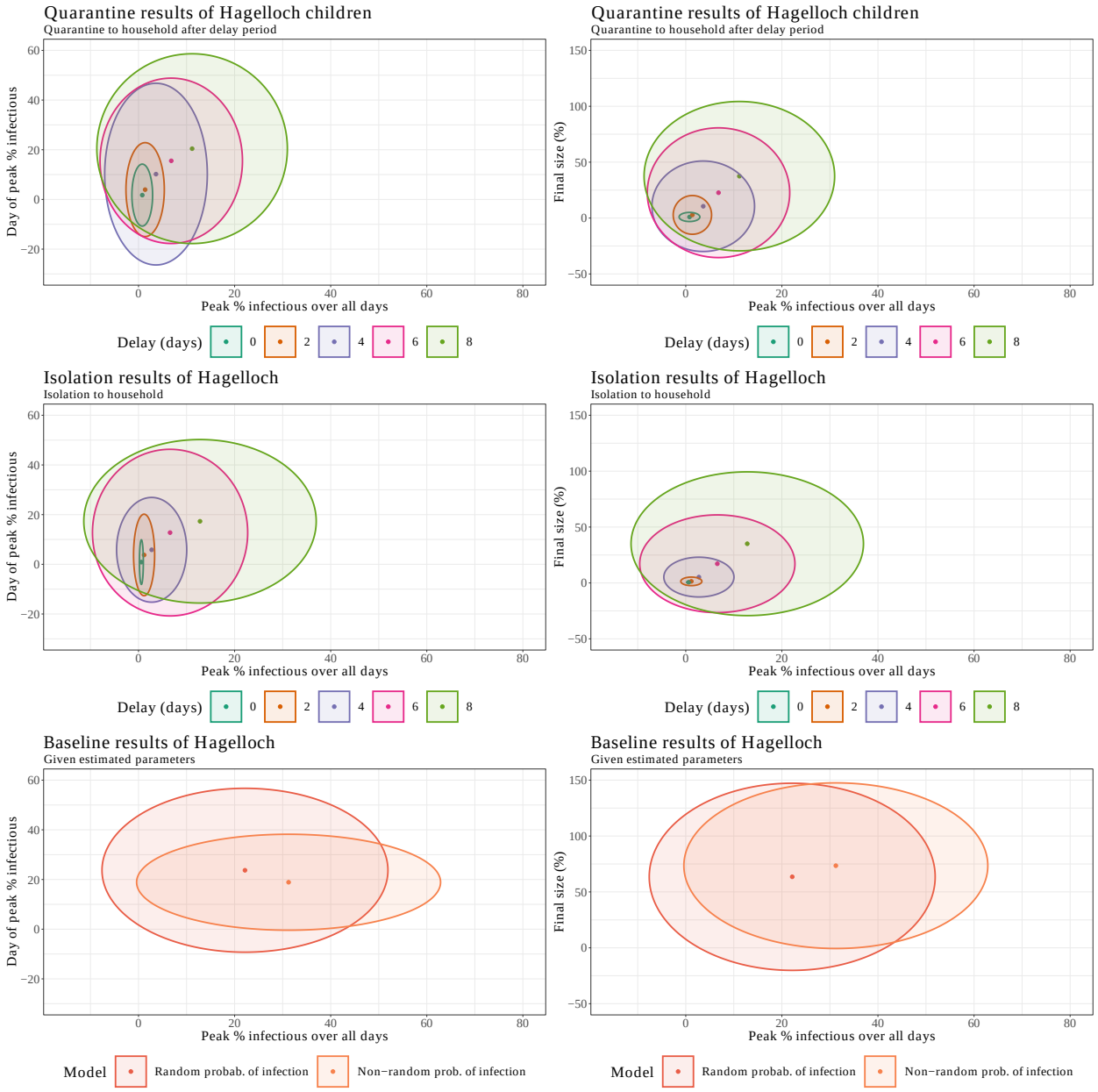

Figure 7.1: Simulation results of isolation and quarantine routines along with baseline simulations for given estimated parameters from Chapter 5. Here, Each each AM was run 100 times with $\hat{\beta}_1 = 0.43, \hat{\beta}_2 = 0.23, \hat{\gamma}_1 = 0.10, \hat{\gamma}_2 = 0.09$.

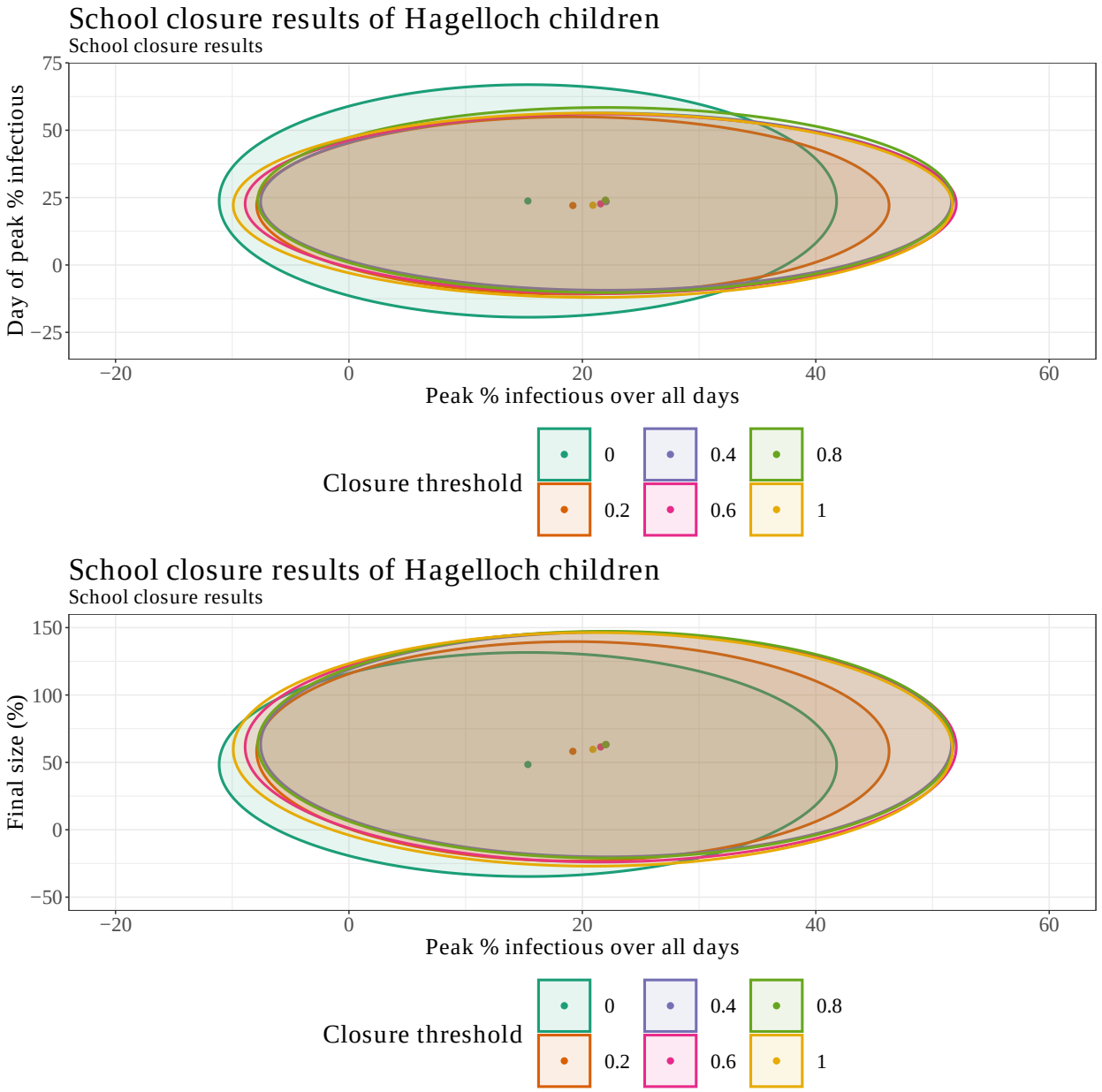

Figure 7.2: AM scenario of school closure for 1st and 2nd class of Hagelloch

have the same results as the "Random probability of infection" model in Figure 7.1 (bottom right). We see that the average estimates for the final size when the threshold is larger than $\rho_c > 0.4$ is actually greater than not closing the school at all. This may be an unintuitive result but 1) we see that the results are highly variable to begin with and 2) Grefenstette et al. (2013) have shown that we can actually prolong an epidemic through school closures.

The results of this simulation do not imply that closing schools is an inviable strategy to prevent the spread of measles, in general. However, it does show how initial calibration can effect our interpretation of the resulting AM. In this scenario, we remove classmates as contacts once the threshold is met, but the remaining population still has an equal chance of interacting with one another. The large variance is especially prevalent in the Hagelloch data set because of relatively small population.

7.5 Chapter summary

We examine two hypothetical scenarios with our stochastic CM-AM pair using parameter estimates fitted to the Hagelloch data in Chapter 5. These two hypothetical scenarios are examples of what policy makers may be interested in when forming responses to potential epidemics. In these scenarios, we examine potential outcomes from imposing quarantine and isolation routines and school closures on our agents.

There are many ways to analyze an epidemic, and we focus on peak infectious %, peak infectious % day, and final size of the epidemics. In order to analyze these hypothetical scenarios, we use our stochastic CM-AM pair as implemented in our `R` package `catalyst`.

For these scenarios with tangible intervention routines, we implement interventions by restricting the number of infectious contacts of each agent when certain conditions are met.

For the first scenario, quarantine and isolation, we are remove almost all contacts except for the housemates after some delay d. Unsurprisingly, we find that isolation is less effective than quarantine and isolation combined. We see that if we isolate infectious agents quickly enough, we can reduce the final size of the epidemic up to 40%!

For school closure, the results are much more inconclusive, having large and often uninformative CIs for our estimates of days of *peak % infectious*, *peak % infectious*, and *final size*. We see that having a closure threshold over 0.4 may result in even larger final size than having not shut down the school at all, but this result is not statistically significant.

For both of these contact restriction scenarios where we directly adjust $\hat{\eta}$ to create local changes within the model, we find that we have very large CIs, many of which, especially in case of school closure, are uninformative. An important reason as to why the CIs are so wide is that the probability that an agent is going to become infectious is now itself a random variable since it is based on $\hat{\eta}$, an issue we will address in the future.

Overall, we see that we can learn much from running hypothetical scenarios with our AM, including how much we would like to reduce the infectivity of a disease and the usefulness of isolation, quarantine, and school closure routines. Moreover, since our AMs are generated using parameters estimated directly from the stochastic CMs, we have more of a reason to trust in our results.

That said, these analyses raise some important issues for modelers. The first issue is that these analyses are conditional on the disease parameter estimates. Future analyses may benefit from putting a prior on these disease parameter estimates to further propagate uncertainty.

A second issue is whether or not to use the observed versus expected number of infectious contacts. On one hand, using the observed number allows our simulations to closer mimic reality. That is, in the simulation, an agent has some chance of receiving an infection from the agents she actually contacted. When we use the expected values, we gloss over these direct connections in favor of an aggregate probability of becoming infectious. However, the results of this are apparent, for example, in Figure 6.1 as compared to Figure 7.1. The CIs for summary statistics for the models using the expected number of infectious contacts compared to the models using the observed number of infectious contacts are much smaller and more informative. This difference in CI width is partly exacerbated by the relatively small number of agents in the simulation. Still, using a random probability of transition, as opposed to a non-random probability of transition, naturally results in larger uncertainty.

A third issue is how interventions in reality may not make as much intuitive sense after they are implemented in the model. For example, we refer to our school closure example where we assumed the agents interacted homogeneously with one another. The result of this is that while closing down the schools did remove contacts from agents, the agents still could interact equally with the remaining population. As such, results and their interpretations only make sense in the context of our selected model.

We finally note that the CM-AM pair here allows us much freedom, as opposed to using a CM. We began with a stochastic CM-AM pair where the CM and AM are equivalent in distribution in terms of the number of agents in the number of states at each given time. However, as soon as we implemented the intervention routines, we *changed* the model. Using the AM allowed us to easily implement these routines, especially using the function $\hat{\eta}(t, n)$, the number of observed infectious contacts of agent n at time t. It is unclear how one would implement these preventions while using the CM in a modular and flexible manner.

In conclusion, we were able to use the strengths of both CMs and AMs in this analysis and as a result can provide recommendations to policy makers and scientists of how we should approach similar outbreaks of measles. Using the CM framework, we were able to perform model selection and estimate parameters on the existing measles data. Once our models were selected, we leveraged the AM framework and our estimated parameters to change the model and agents, locally. This allowed us to analyze important prevention behavior such as reducing the infectivity parameter, isolating and quarantining agents, and finally closing down schools.

The Hagelloch data set provided an ideal test ground for our methods, as it is a feature-rich, yet small in terms of number of actual agents. In the next set of chapters, we will analyze our stochastic CM-AM pairs on a much larger data set in terms of number of agents but smaller in terms of the number of features the agents have.

Chapter 8

Ebola: parameter estimation

While the previous set of chapters explored fitting a stochastic CM-AM pair to a data set from 1861, in this set of chapters we explore a recent outbreak, the Ebola outbreak in Western Africa from late 2013-2015. We limit our analysis to Western District, Sierra Leone, which includes the nation's most populous city and capital, Freetown.

We highlight how the stochastic CM-AM pair is applicable to modern settings. The primary differences between the measles and Ebola case studies are two-fold: 1) the Ebola outbreak encapsulates a much larger region, in terms of geographical area, number of infections, and total population size; and 2) the Ebola person features in the available data are sparse, especially in comparison with the measles outbreak. We show that regardless of the size, the stochastic CM-AM pair is still useful and can be used to investigate relevant questions. In this set of chapters we:

1. Describe the features of Ebola and how this guides our model selection
2. Examine the data for the 2014-2015 Ebola outbreak
3. Fit models and perform model selection on said data
4. Examine AMs with respect to
   (a) the effective population size N
   (b) sensitivity to initial conditions
   (c) basic agent interaction restrictions.

Instead of repeating the analysis done for measles (e.g. looking at reduction in β, isolation and quarantine routines, and school closure), we instead focus on sensitivity of the results to initial conditions in the model. Analysis of the Ebola outbreak is important because of the influx of recent studies of the disease in the

field of epidemiology and because of the even more recent (and more deadly with a mortality rate of 67% in the Democratic Republic of the Congo) outbreak of Ebola in Central/Eastern Africa (Cohen, 2019; Drake, 2019).

We first estimate disease parameters using a stochastic CM. Again we use the SIR disease-states. Here, we concede that having an exposure state (and hence using an SEIR model) may be more appropriate than not having one. However, the data we have does not allow us to easily or accurately infer such a state. Moreover, we find that the SIR disease-states models fit the data well, after adjusting for N, the effective population size.

In this chapter we study the effects of adjusting N, the population size, which was previously considered to be a fixed value. Under the CM framework presented in Chapter 2, all susceptible agents must interact homogeneously with infectious agents, or in other words, have the same probability of becoming infectious from one time step to the next. The validity of the assumption that individuals within a population interact homogeneously with one another is closely associated with the magnitude of N. Moreover, the population size N is seen in both the stochastic K&M Binomial probability of becoming infectious ($p_t = \frac{\beta I(t)}{N}$), as well as in the Reed Frost version ($p_t = 1 - (1 - \beta/N)^{I(t)}$). We now treat N as a parameter of the model. To be clear, we still assume that N does not vary with time. Rather, we now treat N as the *effective* population, i.e. there exists some group of individuals of size N in which the assumptions of the CM hold.

As another focus, we analyze the sensitivity of model results to initial parameters, especially with regards to the location of the initial infection. We compare the disease spreads spatially when we assume homogeneous interaction of agents versus heterogeneous interaction of agents. More specifically, in our simulations, interactions of agents are determined by physical distance among agents.

This chapter proceeds as follows. In Section 8.1, we overview some of the important details of an Ebola infection and discuss the data, especially with in regards how we fit our stochastic CM-AM pair. Following that, in Section 8.2 we fit a stochastic CM-AM pair and estimate parameters to use when analyzing hypothetical scenarios. Finally, in Section 8.3, we summarize the chapter.

8.1 Exploratory Data Analysis

According to Baize et al. (2014), a new strain of *Zaire ebolavirus* (EBOV) was identified as the cause of death of an individual, who is now considered to be "Patient 0," that occurred at the end of 2013 in Guinea. Following the initial case, the disease spread through Western Africa, where over 30,000 probable, suspected, or confirmed cases were reported between 2014-2015. One of the reasons why outbreaks of Ebola are treated very seriously is due to the disease's high fatality rate, which is over 30% in our data set and closer to 70% for the 2019 outbreak in the Democratic Republic of the Congo (DRC) (Drake, 2019).

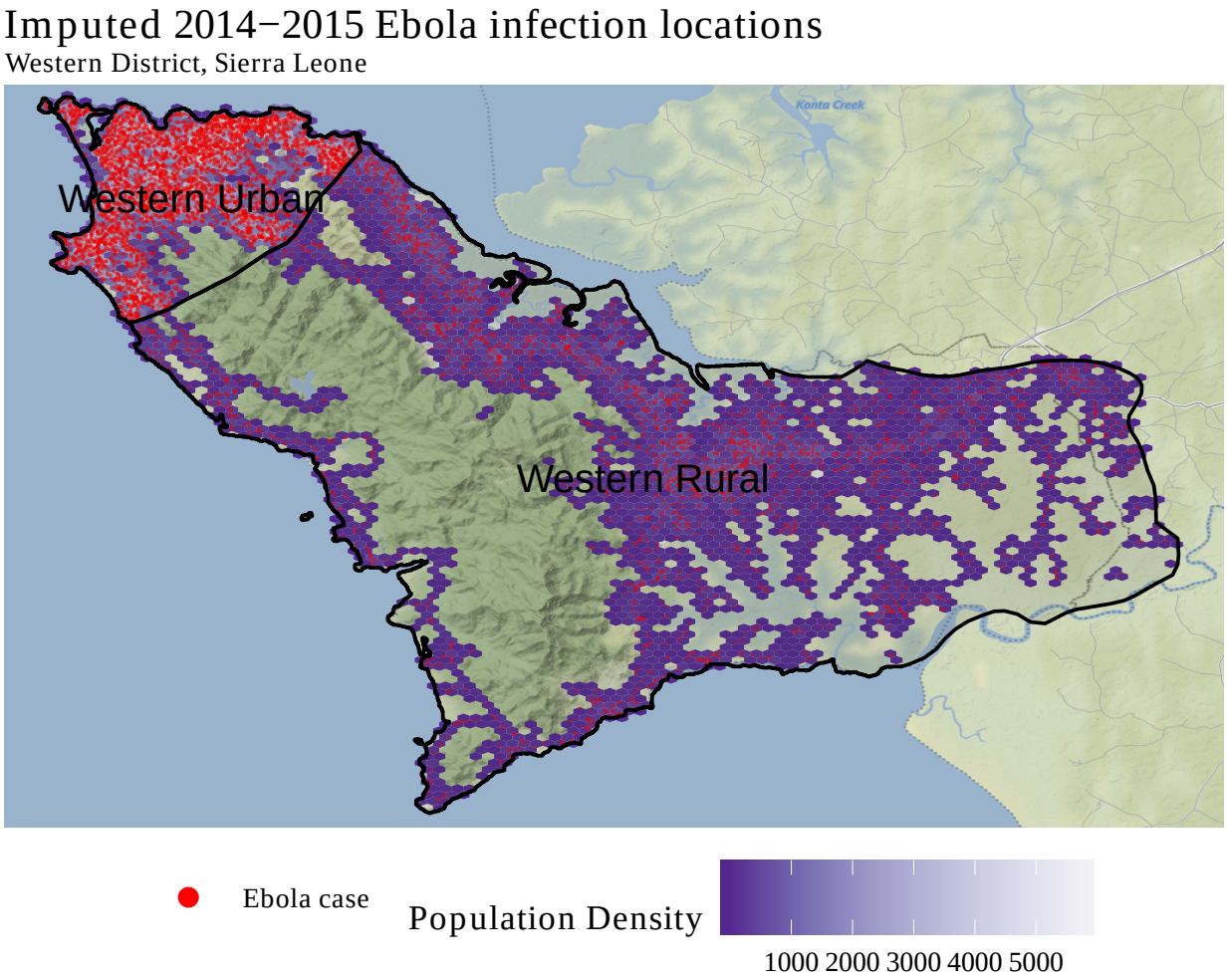

Figure 8.1: Map of Western Urban and Western Rural, Sierra Leone. The North Western part consists of Western Urban (where Freetown is) and the remainder is Western Rural. The population density is plotted according to the synthetic agents produced by SPEW and supplemented further here. The red dots represent imputed infection locations of Ebola between 2014-2015.

Unlike measles, Ebola virus is transferred through direct contact with blood or bodily fluids from a body (alive or dead) infected with the virus or through the consumption of or contact with infected vectors including fruit bats or other primates. Symptoms of Ebola include fever, diarrhea, vomiting, and unexplained hemorrhage and appear 2-21 days after contact with the virus (Centers for Disease Control and Prevention, 2019). A unique feature of Ebola is that the disease can only be spread by an infectious person who has shown symptoms. There are currently no licensed antiviral drugs to treat Ebola and so prevention of the disease is of utmost importance. However, the recent vaccine rVSV-ZEBOV, according to Henao-Restrepo et al. (2017) has a near 100% success rate 10 days after the vaccine has been administered.

Table 8.1: Subset of reported Ebola cases in Western District in Sierra Leone.

Inferred Onset Date	Treatment Center	Final Status	Age
2013-12-26	NA	NA	53
2014-12-27	NA	NA	35
2015-03-26	NA	Dead	NA
2014-11-14	NA	Dead	1
2014-10-23	Freetown 42	Alive	29
2014-12-27	Freetown 42	NA	14

The data, which is available in the supplementary material in Backer and Wallinga (2016), was collected by the World Health Organization (WHO) and ranges from reporting dates of January 2014 through September 2015. The countries infected with Ebola are in Western Africa and include Guinea, Sierra Leone, and Liberia and more. The data consists of over 33,000 confirmed, probable, or suspected cases of Ebola Virus. Of these cases, 21,451 occurred in Sierra Leone. The data are available at the district level, which is equivalent of a US state. Of these cases in Sierra Leone, 8,802 cases were reported in the Western District, which is the area on which we focus in this analysis.

A sample of the data is shown in Table 8.1. As seen in the sample, the data is incomplete. Every record has a date of inferred onset. Of the 8,802 cases, 8,531 do have the age recorded (97%). The treatment center is reported for only 1,459 of the cases (16%). There are 42 treatment centers in Western Urban that were used and 26 centers in Western Rural along with 29 treatment centers outside of Western district. Over 30% of the reported cases resulted in death (2,747 total).

The distribution of the reported ages is shown in Figure 8.2. From the figure, we see that children under 5 years old are the most common group reported. The very young and very old were the most likely to die from Ebola.

Overall, we see that the reported Ebola cases are not very feature rich, especially in comparison to the Hagelloch data set. We do not know any demographic information about the infected people nor location below the district level.

Infamously, the outbreak of Ebola in Guinea began on December 27, 2013 with Patient 0. From there, the disease spread slowly in Western District until July 2014 (approximately 200 days after the first infection). At that point, the disease spread rapidly, peaking around November of that year. The disease then declined, albeit with the occurrence of mini outbreaks through 2015. The spread of the disease is shown in Figure 8.3. We initially assume all $N = 1.4$ million agents (the population of Western District) are initially susceptible, and use the inferred onset date as the day of infection, $t^*_{1,n}$. We impute the maximum time before recovery

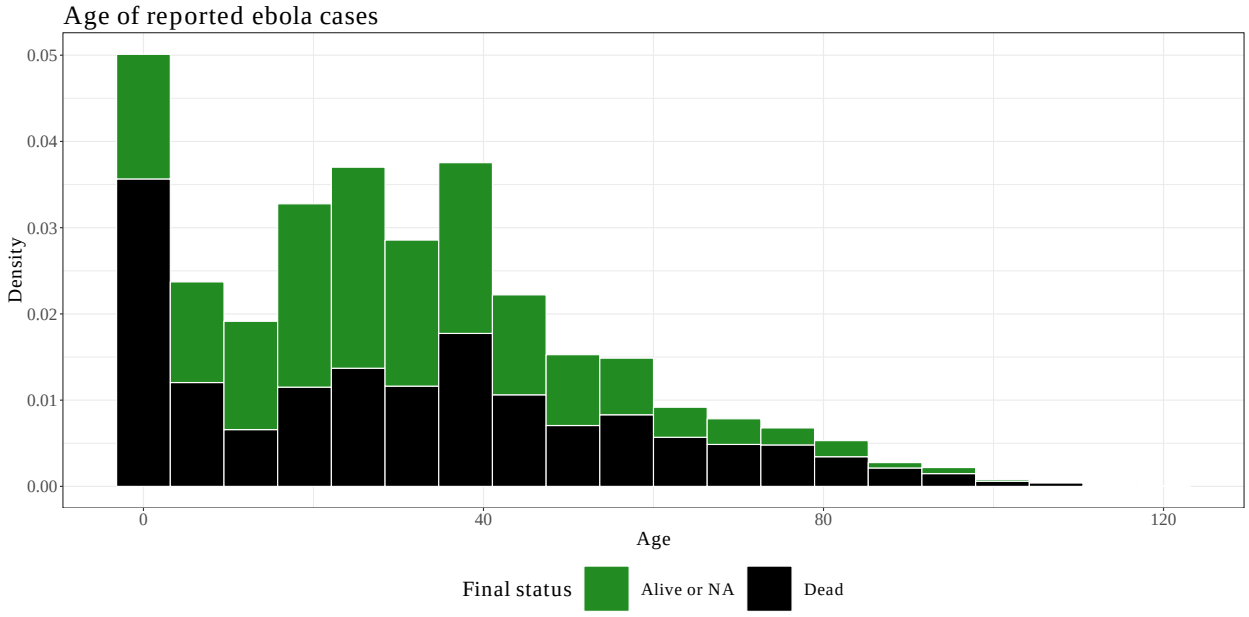

Figure 8.2: Stacked histogram of reported ages, grouped by final status.

for case n, $t^*_{2,n}$ (possibly censored at time $T = 605$),

$$Z_n \overset{iid}{\sim} \text{Poisson}\,(\lambda = 9)$$
$$t^*_{2,n} = \min\left\{t^*_{1,n} + Z_n, T\right\}$$

where $t_{1,n}$ is the maximum time before infection of case n and Z_n is a random Poisson draw with mean $\lambda = 9$. The reason why $\lambda = 9$ is chosen is because Ebola symptoms appear 2 to 21 days after contact with an infectious individual an average of 8 to 10 days (Centers for Disease Control and Prevention, 2019). However, we note the choice of a Poisson random variable is rather arbitrary and is something we will investigate in future work.

8.1.1 Demographics

In 2015, there were approximately 1.4 million people in Western District, which is sub-divided into Western Urban (∼1 million people) and Western Rural (∼0.4 million people) (SPEW, 2017).

The region is shown in Figure 8.1. Geographically, Western Urban's area is about 8 times smaller than Western Rural despite having over twice the amount of people. Moreover, note that a large portion of Western Rural is the Western Area National Park, which is a UNESCO tentative world heritage site (World Heritage Cites, 2018) and is sparsely populated.

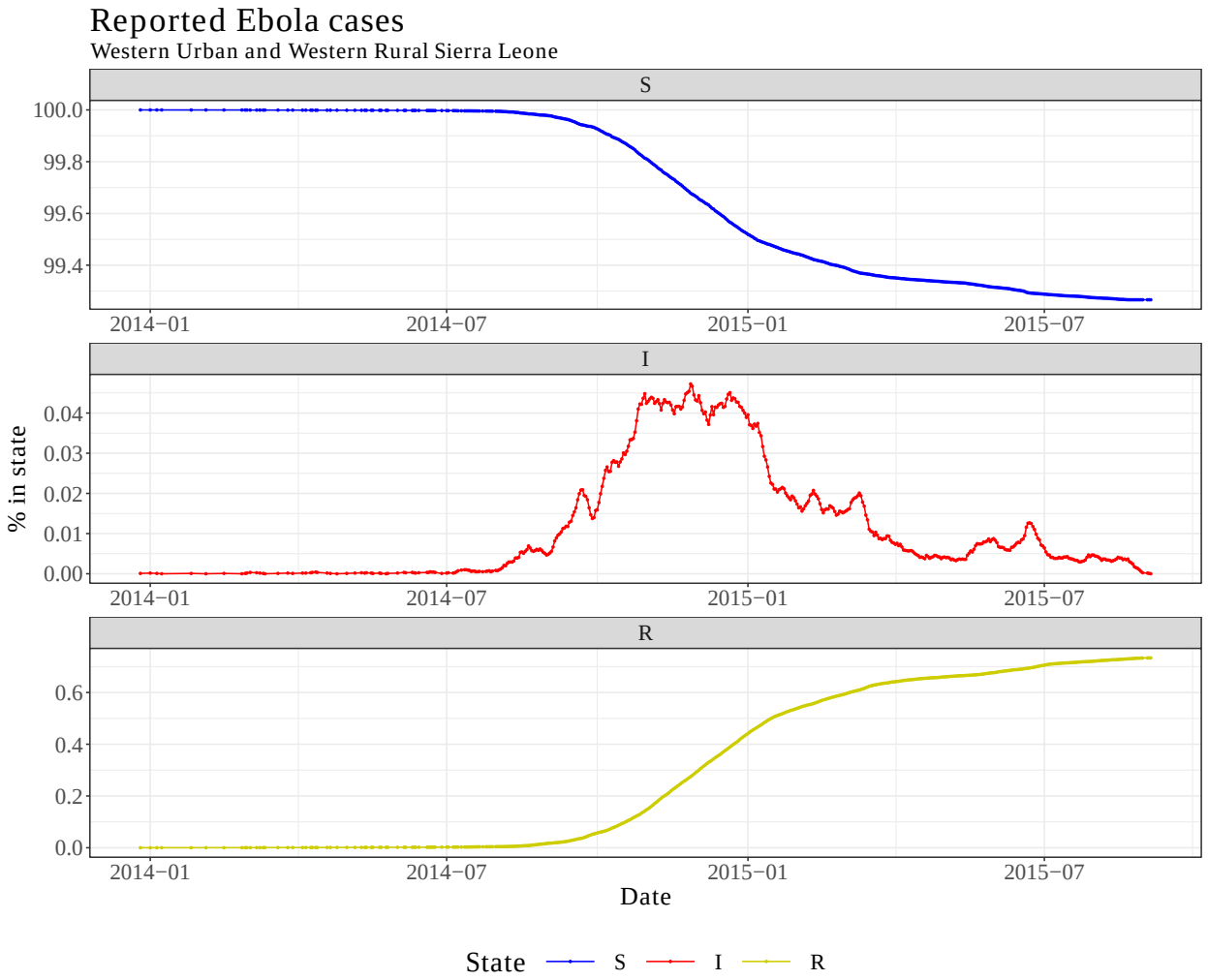

Figure 8.3: Ebola cases in Western Urban and Western Rural Provinces, Sierra Leone. We plot the observed infection dates which have been imputed to SIR format where the recovery time is the infection date plus a Poisson random draw with mean $\lambda = 9$. The susceptible population is taken to be $N = 1.4$ million people.

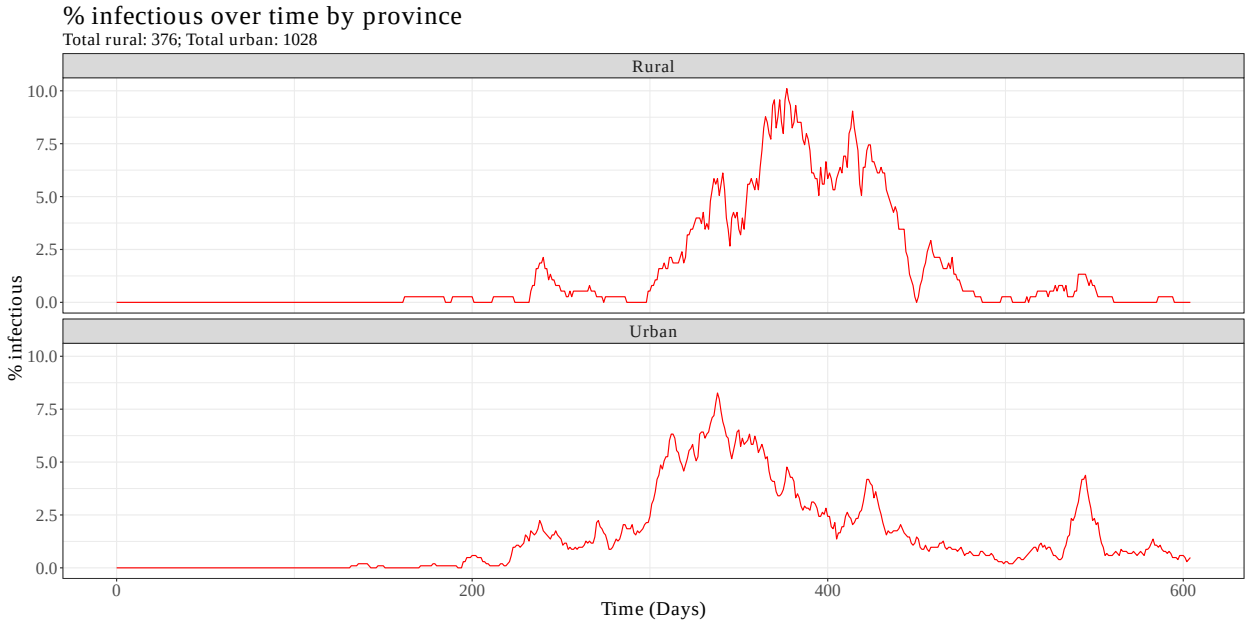

Figure 8.4: Ebola cases in Western district for different age groups. We plot the observed infection dates which have been imputed to SIR format where the recovery time is the infection date plus a Poisson random draw with mean $\lambda = 9$.

In terms of the observed Ebola incidence data, there are two primary demographic features of interest: 1) age and 2) ID of treatment center. Age is reported for all but 3% of the reported Ebola cases. However, only about 16% of the observed Ebola cases have a reported treatment center. The treatment centers are identifiable by district and number but their exact name and geographic location are unknown.

In Figure 8.4 we plot the incidence as a percentage over time grouped by whether the treatment center the case attended was in Western Urban or Western Rural. From this figure, it seems that incidence moves from the urban to the rural areas. In Figure 8.5, we plot the incidence as a percentage over time grouped by age categories: 0-5, 5-15, 15-30, 30-45, 45-60, and 60+ years old. When we say percentage, we mean out of the total number of individuals in that group. One trend of note is that the percent of infectious over time for the 0-5 year olds looks more similar to the 60+ year olds than to the other groups. We also see that the 0-5 year olds and 60+ year olds have at least two distinct peaks whereas the other age groups (besides the NA category) have only one distinct peak.

In short, there are very few demographic features to go along with the Ebola case inferred infection dates. In comparison to the Hagelloch measles data set, we are missing sex, household structure, household location, class, and purported infector ID.

Since the Ebola data set is poor with respect to demographic features of the infected agents and says nothing at all about the other susceptible agents, we supplement the Ebola incidence with a synthetic agent data set. Specifically, we use the Synthetic Population and Ecosystem of the World (SPEW) synthetic

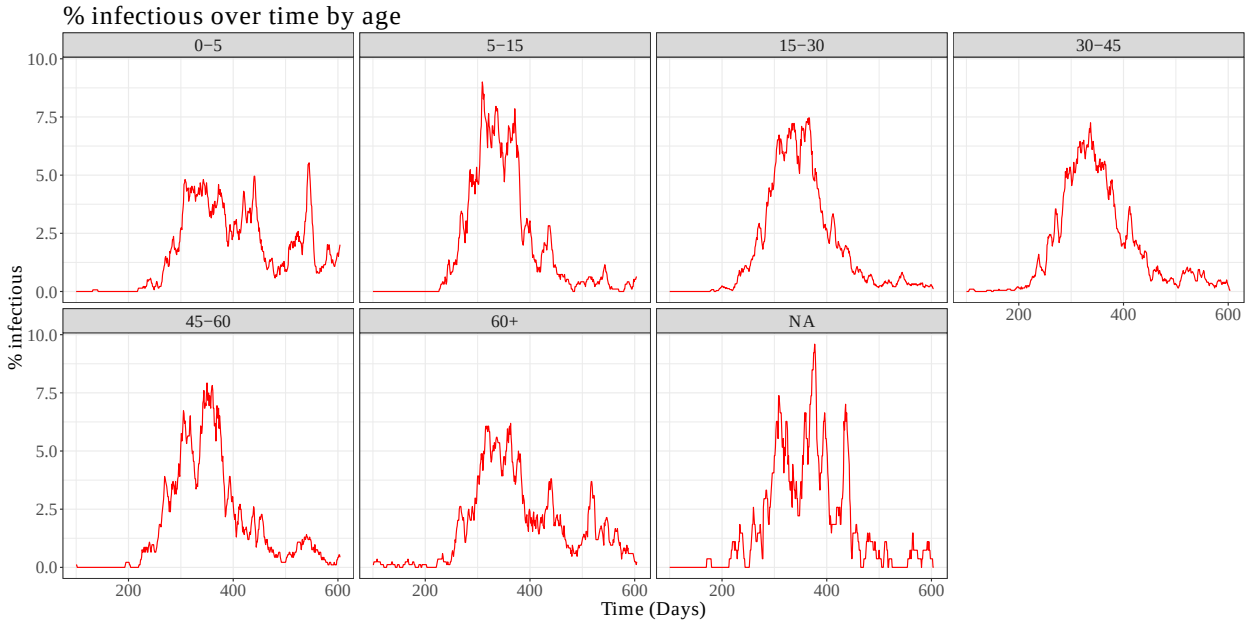

Figure 8.5: Ebola cases in Western Urban and Western Rural Province treatment centers, Sierra Leone. We plot the observed infection dates which have been imputed to SIR format where the recovery time is the infection date plus a Poisson random draw with mean $\lambda = 9$.

agents produced by Gallagher et al. (2018) for Sierra Leone. The agents are generated so as to best maintain demographic distributions and household structure within a region. We further supplement the SPEW synthetic agents with road-based sampling which follows the logic that people tend to live near roads. The resulting SPEW population is shown on the map in Figure 8.1.

Finally, we combine our Ebola incidence data set with the Western District synthetic agents. For each incidence case, we examine the corresponding synthetic agents who are plus or minus 2 years of age of the reported case. Of those agents in the proper age range, we uniformly at random select one of them to be the corresponding demographic characteristics for the incidence case. Once an agent is selected to correspond to an incidence case, it is removed from the set agents to be sampled from for the remaining incidence cases.

The main benefits of supplementing the Ebola incidence cases with SPEW agents are two-fold: (1) we gain a household structure for our agents, and (2) we have a much finer spatial granularity of our cases, which will be useful once we begin to consider heterogeneous interactions of our agents.

Unfortunately, the locations of the Ebola incidence cases are drawn independently from any location data (because we have none or very little of it), meaning that we will not see any obvious pattern in spatial transmission of the disease over time in our imputed data set of incidence cases. However, we still consider this a useful exercise because we will explore how the spatial features of the infectious cases change as a result of varying interaction conditions for the agents based on their proximity to one another.

8.2 Ebola model selection

The first choice we have to make to structure our stochastic CM-AM pair is to fix our disease-level states. Once again we use the SIR disease-level states, although we do note there are some concerns with choice. One notable feature about Ebola is that the disease can be transmitted even when the victim is deceased, typically through interactions with the body via funerals or at hospitals Drake et al. (2015). However, the data to which we have access does not provide any insight about these transmission states, and so like in the case of measles case study, we use the SIR disease states.

Another issue one may be concerned with is birth and death in the population since the epidemic occurs over a period of two full years with a population of 1.4 million people. In the literature, this problem is commonly dealt with by introducing birth and (non-Ebola) death into the population such that the final population remains constant N (Anderson and May, 1992; Allen, 1994; Zaman et al., 2009). We assert that since less than 1% of the population of the Western district is infected over the course of two years, adding in birth and death unnecessarily complicates the model, especially given that we do not know *how* the population changes over time.

For the measles outbreak in Chapters 5-7, we focused on partitioning the agents into different groups, but in this chapter we focus on the total population N used in the model, which until this point has been treated as a fixed value.

Recall the K&M deterministic CM-SIR difference equations are the following,

$$\begin{cases} \frac{\Delta S}{\Delta t} &= -S \times \beta \frac{I}{N} \\ \frac{\Delta I}{\Delta t} &= S \times \beta \frac{I}{N} - I \times \gamma \\ \frac{\Delta R}{\Delta t} &= I \times \gamma \end{cases}$$

The role N plays in these equations is the number of individuals in the total population. Moreover, the model specification implies that all N of these agents interact homogeneously with the infectious agents. In the Hagelloch measles outbreak, we had a strong argument that N was equal to the final size of the disease and also explored what happens to $\mathcal{R}_0$ when we considered the full population size (adults and other immune individuals). For the measles outbreak, even when considering the larger N, it was still on the same order of magnitude as the final size. Additionally, it is not unreasonable to assume that the population interacts approximately homogeneously since Hagelloch in 1861 was a small and isolated village.

Compare this to the Ebola outbreak in Western District Sierra Leone, where only 8802 cases were reported out of a population size of 1.4 million people, less than 1% of the population. It is difficult to imagine any scenario where 1.4 million people interact homogeneously. At the same time, it is also difficult to imagine that the 8802 reported cases only interacted among themselves. Moreover, Western District is split into an

urban and rural region (see the map in Figure 8.1), and it is highly unlikely that a person living in a small town in Western Rural would interact in the same manner with an Ebola-infectious person who is located in the more urban Freetown.

We therefore propose that there is some sub-population of size N where the susceptible people interact, approximately, homogeneously with the infectious individuals. We explore this scenario by treating N as a model parameter, in addition to β and γ. We do this by first finding model parameters by minimizing the joint mean square errors, and in future work, we will consider likelihood based methods where we consider N to have some probability distribution.

As we are looking for parameter estimates for our AM, we find it reasonable to first explore fitting a deterministic SIR-CM as a function of $f_i(\beta, \gamma, N)$ (with values scaled between 0 and 1 (separately) for each of the S, I, and R values), especially since the expected value of our Binomial SIR model is unbiased.

Specifically, we find estimates of β, γ, and N,

$$(\hat{\beta}, \hat{\gamma}, \hat{N}) = \arg\min_{\beta,\gamma,N} \frac{1}{T} \sum_{i=1}^{3} (\mathbf{x}_i - f_i(\beta, \gamma, N))^2 \tag{8.1}$$

where $\mathbf{x}_i$ are the observed SIR Ebola data that are first scaled between 0 and 1. Note that the number of susceptibles in the *data* changes as N changes. The results of this are shown in Table 8.2 for different values of N, specifically when the agents are assumed to interact homogeneously (N= 8802), the "best" N (N = 18758), $N = 10^5$ an intermediate value, and $N = 1.4 \times 10^6$, the actual population of the Western District. The best $\hat{N}$ value is found to be $N = 18758$ using Nelder-Mead optimization, minimizing the objective function in Eq. (8.1). Using this, we find $\hat{\beta} = 0.159$ and $\hat{\gamma} = 0.122$ where consequently $\hat{\mathcal{R}}_0 = 1.31$ (95% CI: [1.21, 1.40]). In contrast, when $N = 8802$, $\hat{\mathcal{R}}_0 = 1.63$ (95% CI: [1.44, 1.82]), and when $N = 1.4 \times 10^6$, $\hat{\mathcal{R}}_0 = 1.05$ (95% CI: [1.05, 1.05]). These results show that N provides a very important role in estimating $\mathcal{R}_0$ both in terms of the value of $\mathcal{R}_0$ but also the widths of the corresponding CIs, and at the very least, this analysis provides an upper and lower bound on $\mathcal{R}_0$, somewhere in between 1.05 and 1.82.

Table 8.2: Joint Mean Square Error (MSE) for observed vs. fitted SIR models with varying N along with an estimate of $\mathcal{R}_0$ and a 95% CI interval. The highlighted row is the model with the minimum MSE for all N, β, and γ.

N	MSE	$\hat{\mathcal{R}}_0$	Lower bound	Upper bound
8802	0.10	1.63	1.44	1.82
18758	**0.01**	**1.31**	**1.21**	**1.40**
100000	0.13	1.05	1.02	1.07
1400000	0.69	1.05	1.05	1.05

We take a closer look at the best model according to the minimum joint MSE. We plot the model and observations as a zoomed in ternary plot in Figure 8.6. The observations are plotted in black and the

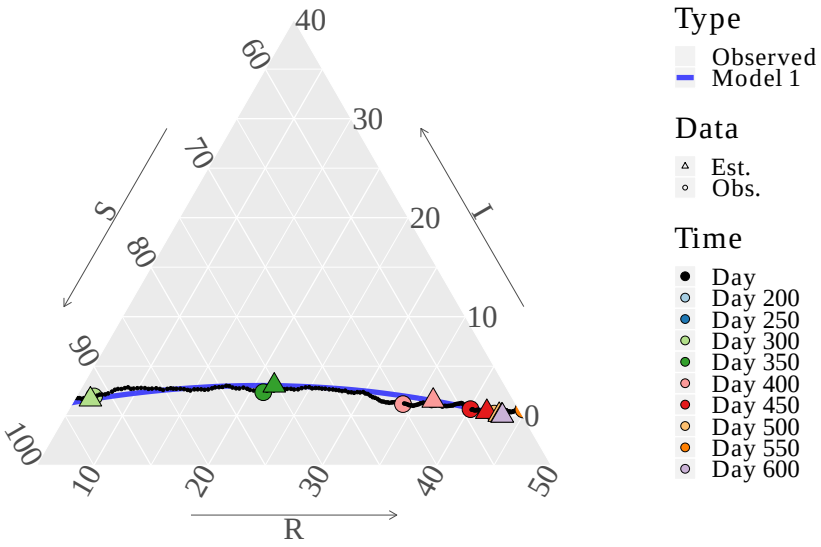

Figure 8.6: Observed Ebola SIR data from 2014-2015 for Western District Sierra Leone and best fit SIR model with $\beta = 0.16$, $\gamma = 0.12$ and $N = 18768$ as a ternary plot.

estimated trend in blue. To show some sense of the time scale, we plot different values of days in different colors, triangles for the estimated values, and circles for the observed values. Overall, we see that, visually the best estimated model fits the data very well, even when taking into account the time scale.

The conclusion we can draw from this model fitting analysis is that the total size of the homogeneous population we assume (or estimate) in the SIR-CM models (both deterministic or stochastic) is extremely important in inferring information about the infectivity and recovery period of a disease.

In terms of our stochastic CM-AM pairs this is important in not only estimating disease parameters to initialize our AM but also about the general dependency structure of the agents. More specifically, the best-fit SIR tells us there is about a community of 18,750 people that have approximate homogeneous interactions with the infectious agents. If we knew more about the infectious agents, e.g. their hospital locations, their families, and their homes along with information about hospital workers, we could use this number to simulate an agent-based model among neighbors.

From this modeling, we also have evidence that the Western District is independent from outside infectors, with perhaps an exception for the initial infection, which supports the results found in Backer and Wallinga (2016). That is, since our best fit SIR model is a very good fit to the data, we do not need to rely on outside actors to sustain the Ebola epidemic. From that, we can instead focus on containing the disease locally instead of having to focus most of the preventive routines for long distance travel of cases.

Finally, estimating N is important in that it greatly improves what we can study with a flexible CM-AM pair. Since AM run time is based on the interaction between agents, it is much more computationally efficient to look at a closed group of 18,750 people rather than all 1.4 million individuals at once. If using the full population of 1.4 million people, we are limited in what interventions we can analyze in our simulations simply due to practical limitations in computing.

8.3 Chapter summary

Ebola is a deadly disease and so it is of utmost importance to better understand its spread through Western District in Sierra Leone, which consists of over 8,000 of the over 30,000 reported cases in the Ebola epidemic of 2014-2015 in Western Africa.

In this chapter we explored the Ebola data set presented in Backer and Wallinga (2016) and estimated the disease parameters corresponding to an SIR model. In particular, we introduced the issue of treating N as a parameter, the effective population size, rather than a fixed value.

In our exploratory data analysis in Section 8.1 we impute recovery dates for reported infection cases based on past CDC information about Ebola. Using this estimate, we find that the Ebola SIR curves look like noisy versions of deterministic SIR curves, especially from July 2014 through 2015, although there looks to be a mini outbreak in May 2015. As such, we find it appropriate to estimate parameters from SIR-CM models. Since Western District is split into Western Urban and Western Rural, we examine the infectious curve in the reported treatment centers in the two regions (although the treatment centers were reported for only 16% of the data). From 8.4 we see that urban infections generally seems to precede rural infections, which may suggest that the infection spreads from the city into more rural regions. We also examine the spread of disease among different age groups (see Figure 8.5) and find the very young and very old seem to have more similar infection curves than those of other ages.

For the model selection, we decide to treat the population as one group with the same β and γ parameters. However, we find that the effective population size N is very important in fitting a SIR-CM to the Ebola data. Specifically, we find that $N = 18,750$ is the optimal number of people according to minimizing the joint MSE of the observed data. The best fit model is shown in the ternary plot in Figure 8.6, and our best estimate for $\hat{\beta} = 0.159$ and $\hat{\gamma} = 0.122$ which indicates a reproduction number estimate of $\hat{\mathcal{R}}_0 = 1.31$ (95% CI: [1.21, 1.40]). We find that the effective population size N is very important when fitting a model and can help guide our AM in terms of the number of people needed to adequately model the population.

In the next chapter, we will explore AM simulations using our parameter estimations from this chapter as a guide. We will examine spatial spread of the disease, the effect of N on our simulations, and sensitivity to initial conditions of our model.

Chapter 9

Ebola: hypothetical scenarios

9.1 Chapter goals

In this chapter, we use the model selection and parameter estimates from Chapter 8 to initialize our stochastic CM-AM pair and analyze how variation in initial parameters effects the results. As AMs are used to analyze both spatial and temporal aspects of an outbreak, we examine both aspects here. Unfortunately, we do not know how our infected cases in the Ebola data are related to one another or where they are located within Western District. Moreover, we only know the treatment center of 16% of the infected cases. As such, it is difficult to say anything conclusive about how the disease spreads spatially.

Instead, we will make simple assumptions about how Ebola spreads throughout the community, namely by assigning contacts to agents based on their household location. We do not claim that this is how the disease spreads in reality but believe it is a good basis from which to guide future AMs when there is more information about how the disease is transferred from person to person. Thus, many of the results in this chapter represent a large-scale proof of concept, in that a CM-AM pair can be used to analyze a region even when N is large.

In this chapter, we explore the following three initial values of the AM and how they influence the resulting spread of Ebola

1. Homogeneous versus heterogeneous interaction of agents
2. Sensitivity to the initial location of infectors
3. Effective population size and number of neighbors.

In Section 9.2, we study homogeneous interaction versus heterogeneous agent interaction. We examine our best fit CM-AM pair, which has homogeneous interaction of agents. The results of this study show

this model fits the data very well, especially compared to any heterogeneous interaction CM-AM pairs with the same population size. However, when using a heterogeneous interaction AM, the spatial spread of the disease differs radically. We demonstrate how the transmission differs and show this statistically.

Following that, in Section 9.3 we examine the sensitivity of the CM-AM pair to the initial infector locations. Since Western District is split into Western Urban and Western Rural we explore the spatial spread of the disease for 1) when all the initial infections are located in Western Urban and 2) when all the initial infections are located in Western Rural. Again, we assume that the heterogeneous interaction of agents is based only upon their physical location and so this is more for illustrative purposes and future CM-AM pairs than for any conclusions regarding this specific Ebola outbreak.

Then in Section 9.4 we examine both the effective size of the population, N, along with the limiting the number of neighbors, M. We demonstrate how variation in N and M effect summary statistics of the spread of a disease, and we show the differences in computational memory and time needed to carry out such simulations.

Finally, in Section 9.5, we summarize and highlight the results found in this chapter.

9.2 Homogeneous versus heterogeneous agent interaction

In this section, we compare homogeneous interaction of agents versus interaction of agents based on their household location. In Chapter 8, we found that the best fit SIR-CM has the following parameters: effective population size $\hat{N} \approx 19000$, $\hat{\beta} = 0.159$ and $\hat{\gamma} = 0.127$ and homogeneous interaction of individuals. We propose that it is reasonable to assume there exists some group of about 19000 agents that act approximately homogeneously with the infectious agents, and this is supported by the best-fit model's low MSE and visual diagnostics via a ternary plot. However, it is a whole other and quite more difficult question to ask *which* agents belong to such a group. If the data contained more demographic information about the infected cases and where they were treated we would likely to be state something more verifiable. Since we do not have this information, we instead demonstrate possible scenarios of agent interaction and determine how this effects the spread of the disease spatially. In this set of simulations, we perform the following experiments:

1. Fix the effective population size N and use estimates of $\hat{\beta}$ and $\hat{\gamma}$
2. Assign appropriate number of agents to be initially infectious
3. Simulate the AM under the assumption of
   (a) homogeneous agent interaction
   (b) heterogeneous agent interaction.
4. Compare the aggregate disease state totals

5. Examine the spatial spread of the disease.

We begin with $N \approx 19000$, which is the best fit N and the corresponding best estimates $\hat{\beta} = 0.159$ and $\hat{\gamma} = 0.127$. To select the agents we randomly sample $P = N/(\text{ave. household size})$ from the 1.4 million available agents, where the average household size is about 8 in Western District. We then union all the household members to this set since household members presumably have ample contact with one another. Of these N resulting agents, we randomly select 11 to be initially infected and 32 to be initially recovered which corresponds to the data on day $t = 200$, right before the number of infectious begin to increase exponentially. These 11 initially infected and 32 recovered are the same for all the following simulations.

Once the initial agents and parameters are determined, we run the stochastic CM-AM pair for $L = 100$ runs. For homogeneous interaction, we assume everyone is a contact of each other and thus has equal chance of infection. The results of homogeneous interaction are displayed in Figure 9.1. Again we see that the CM-AM pair simulations fit the observed data well, perhaps with the exception around day 550 where we may have a mini occurrence of an outbreak. The difference between the model shown here and the one used in the ternary plot in Figure 8.6 is that we use the *observed* number of infectious individuals at the previous time to estimate the probability of a susceptible becoming infectious as opposed to the expected number. The result is the same mean SIR curves but with larger CIs, especially around the peak infectious time.

The estimated final size is 33% (95% CI: [10, 57]%). We estimate the expected duration of the epidemic as 533 days (95% CI: [201, 737]). We estimate the estimated peak size as 2.17% (95% CI: [0.53, 3.80]%) and the day of estimated peak as day 350 (95% CI: [230, 471]). The main take away is that even taking our estimate of the disease parameters to be fixed, the resulting AM produces highly variable results in terms of the summary statistics reported above. That said, at least compared to the Hagelloch measles results in Chapter 7, our CIs do not span the entire space of possible results. The primary reason for the relatively smaller CIs is the larger sample size of agents and secondarily the fact that we are using one estimate of infections parameters for all agents as opposed to separating them into groups.

With respect to the spatial spread of the disease, if the agents mix homogeneously, then we do not expect that the spread of the disease to be dependent on the agents' locations. That is exactly the result we see here. In Figure 9.2 we plot a map of the region of infectious individuals over all $L = 100$ trials. The hexagons are colored by taking the average time of infection of all the agents in the geographical region of the hexagon. Since the color of the hexagons is approximately uniform across the entire region, then the spread of the disease is independent from the geographical location of the agents. The initial infectors at time $t = 200$ are plotted as circles in the map. The initial infectors were assigned randomly from the set of agents (recall that the northwestern part of the map is more densely populated and hence there are more initial infectors in that region).

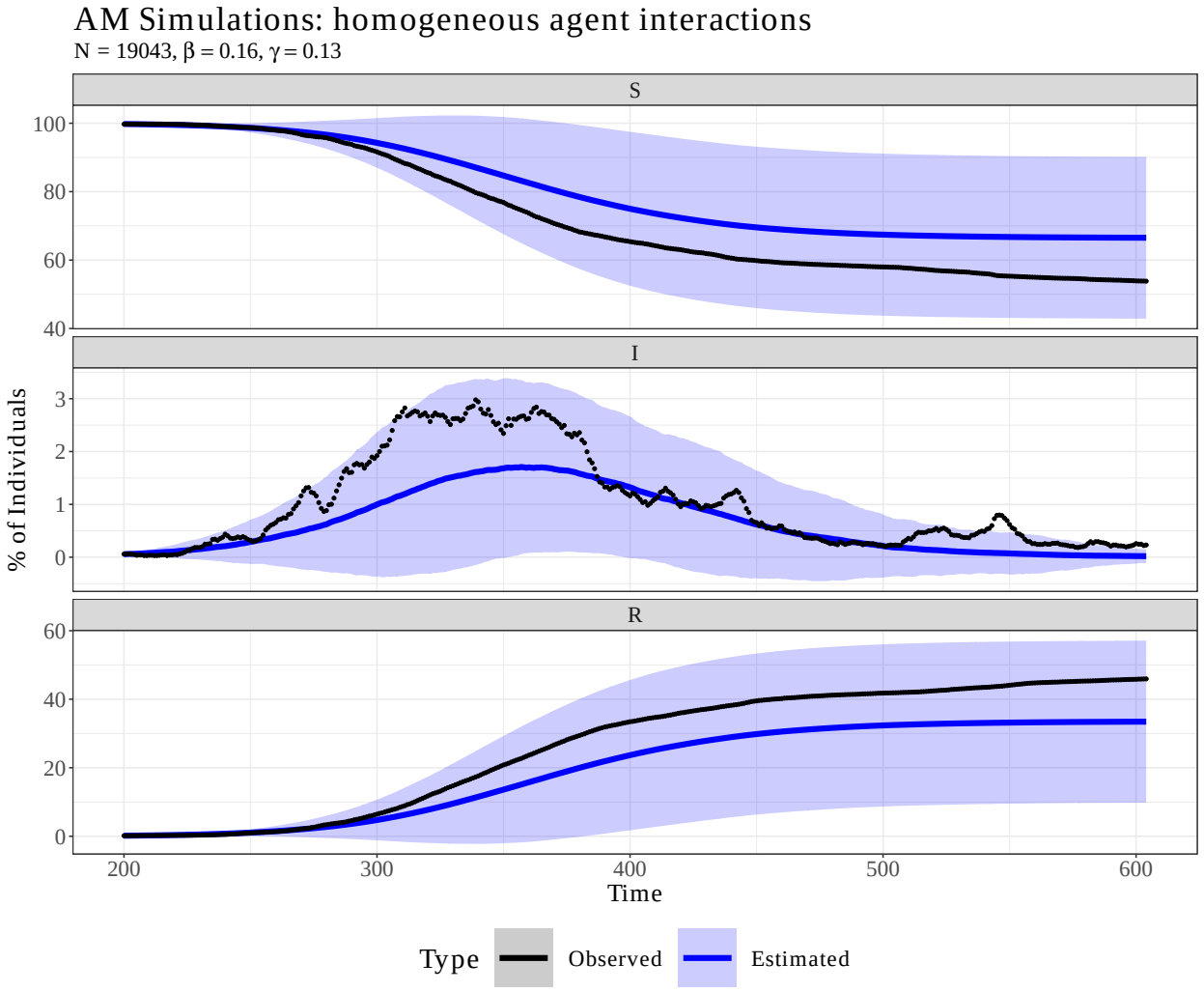

Figure 9.1: SIR curves and 95% CIs for the results of the AM for homogeneous interaction of agents for the best fit model SIR-CM.

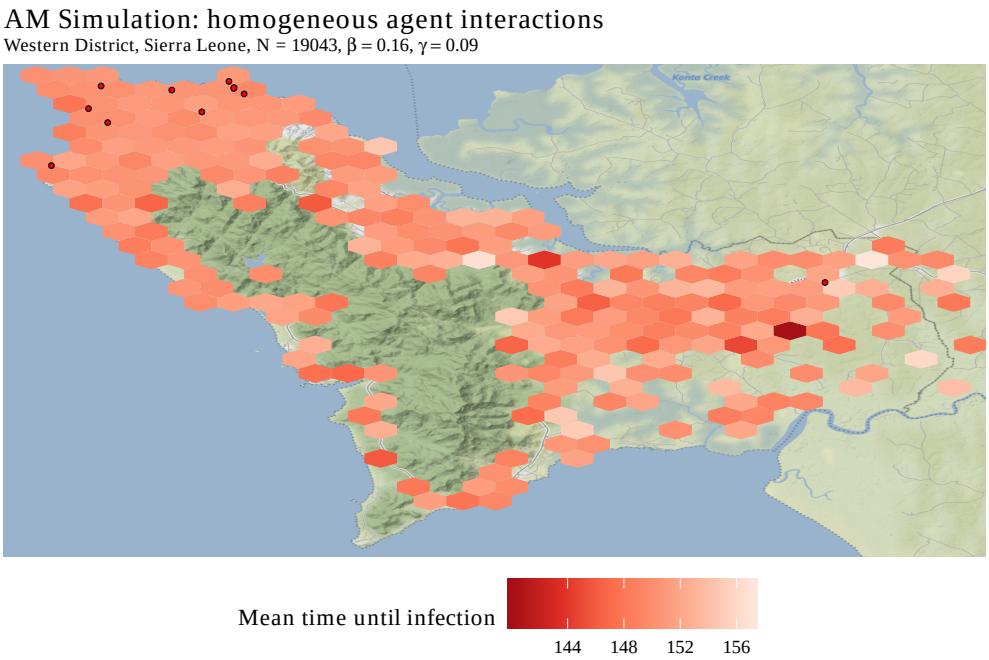

Figure 9.2: Map of infectious agents over $L = 100$ runs where the hexagons are colored by the average time of infection of the agents over all the trials for the results of the AM with homogeneous interaction of agents for the best fit model SIR-CM. The initial infections at time $t_0 = 200$ are plotted as circles.

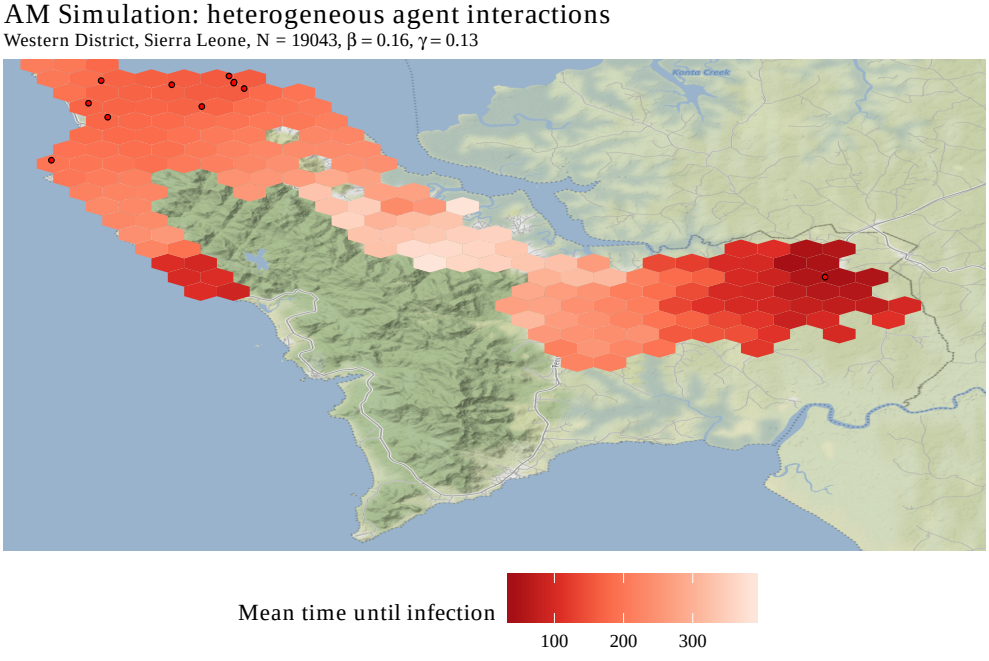

Figure 9.3: Map of infectious agents over $L = 100$ runs where the hexagons are colored by the average time of infection of the agents over all the trials for the results of the AM with simple heterogeneous interaction of agents for the best fit model SIR-CM. The initial infections at time $t_0 = 200$ are plotted as circles.

We summarize the variation in distance via the following statistic, which captures how far away infected individuals are from the initial infectors. For the combined simulations $\ell = 1, \ldots, L$, let J_0 be the indices of the initial infectors and let J_{TL} be the set of indices all of the infections over all the trials. We first find the set that consists of the minimum distance, δ_x between point x in set J_{TL} and point y in set J_0. Our statistic, m is then the median distance in that set. The distance between points x and y, $d(x, y)$ is taken to be the haversine distance. The mathematical formulation of the statistic is shown below in Equation 9.1,

$$\begin{aligned} \delta_x &= \min \{d(x, y);\ y \in J_0\} \\ m &= \text{median} \{\delta_x;\ x \in J_{TL}\}. \end{aligned} \tag{9.1}$$

We find the median distance to be 1.24 miles ($Q_{2.5} = 0.19$, $Q_{97.5} = 2.65$), and the empirical distribution of the minimum distance has a long, thin right tail.

We also look at the heterogeneous interacting agents. In this basic heterogeneous interaction experiment, for $N = 19000$, agents were randomly assigned up to 100 neighbors who were within a one mile radius of the reference agent. The probability of a susceptible becoming infectious from one time step to the next is now

$$p_{t,n} = \frac{\beta \times \#\text{Infectious neighbors of agent n at time } t}{\#\text{ Total neighbors of agent n}}$$

We found that this model, regardless of the disease parameters $\hat{\beta}$ and $\hat{\gamma}$, is a poor fit to the observed data. The reason for this is that the number of neighbors is too small to sustain the spread of the disease. We will examine this more in detail in Section 9.4.

We still find it useful, however, to look at the spatial distribution of the disease for illustrative purposes. A map of the spatial distribution of the disease is shown in Figure 9.3, again where the hexagons are colored by taking the average time of infection of all the agents in the geographical region of the hexagon. We see a clear, continuous change of the gradient of the colors, as smaller times until infection are close to the initial infectors and larger times to infection are farther away from the initial infectors. We see clearly how the disease spreads in Western Urban quickly and into Western Rural. For the heterogeneous interactions, the median minimum distance is 0.93 miles ($Q_{2.5} = 0.19$, $Q_{97.5} = 2.65$).

Finally, the minimum distances are plotted vs. the time until infection for both the homogeneous and heterogeneous interaction of agents in Figure 9.4 with a Loess smoother trend line plotted on top. These scatter plots show that the minimum distance and time until infection are different depending on how the agents interact with one another.

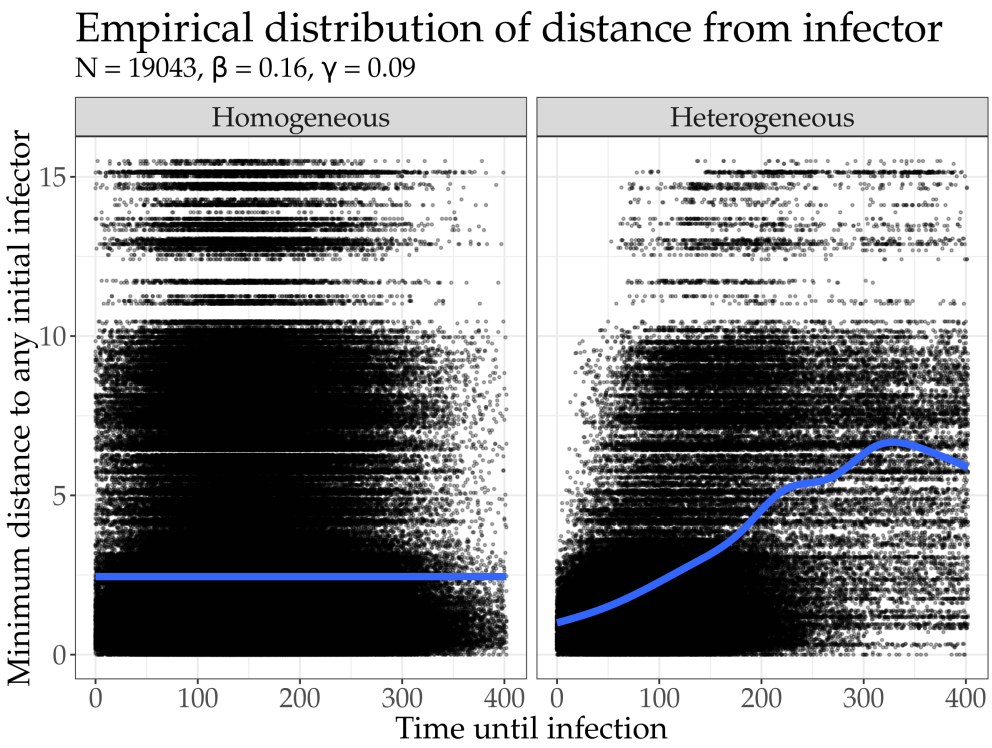

Figure 9.4: Scatter plots of the empirical δ_x from Eq. (9.1) vs. time until infection for the heterogeneous interaction and homogeneous interaction of agents with a Loess smoother trend line on top.

9.3 Sensitivity to initial infection locations

The next aspect we examine is the sensitivity of the CM-AM pair to initial infection locations. Since our treatment centers in the data are split into Western Urban and Western Rural, we examine initial infections located in exclusively in one of these regions. For an AM with homogeneous interaction of agents, the initial locations do not matter since the location is independent of the next infection and so we do not simulate any homogeneous agent interaction CM-AM pairs here. For the sake of comparison, we analyze a CM-AM pair with $N \approx 19000$ agents with $\hat{\beta} = 0.159$ and $\hat{\gamma} = 0.127$ and the heterogeneous interaction scheme outlined above.

Our agents for the Western District are naturally partitioned into Western Urban and Western Rural (see Figure 8.1) where Western Urban has about 2.5 times the number of people than Western Rural, but Western Rural has about 8 times more land area. Under our basic heterogeneous agent interaction scheme of only having the possibility of 100 people from the effective population size N within a one mile radius of the reference agent, we may expect the disease to spread differently based upon the initial infectors.

To analyze the spread of Ebola in Western Urban vs. Western Rural, we perform the following experiment where we

1. Set the effective population size N and randomly sample $N/8$ agents from the SPEW agents and additionally add the household members of those sampled agents to the set of sampled agents
2. Assign each agent neighbor interactions by randomly choosing up to 100 agents within a one mile radius of the reference agent
3. Randomly sample initial infection agents who belong to Western Urban (Rural).
4. Simulate the spread of disease for L iterations using the same initial parameters
5. Compare the results for the initial infections in Western Urban vs. Western Rural.

For our experiment, we set $N \approx 19,000$ and $L = 100$ runs.

For the experiment with initial infectors exclusively located in Western Urban, we estimate the final size to be 21% (95% CI: [1, 42]%), the epidemic duration to be 546 days (95% CI: [378, 764]), the peak infection size to be 1.1% (95% CI: [0, 2.22]%), and the day of peak infection to be 371 days (95% CI: [178, 753]). Moreover, the median minimum distance to an initial infector is 0.77 miles ($Q_{2.75} = 0.10$, $Q_{97.5} = 2.55$).

In comparison, for the with initial infectors exclusively located in Western Rural, we estimate the final size to be 2.94% (95% CI: [0.00, 10.38]%), the epidemic duration to be 380 days (95% CI: [200, 585]), the peak infection size to be 0.24% (95% CI: [0, 0.69]%), and the day of peak infection to be 271 days (95% CI: [200, 439]). The median minimum distance to an initial infector is 1.39 miles ($Q_{2.75} = 0.00$, $Q_{97.5} = 0.25$).

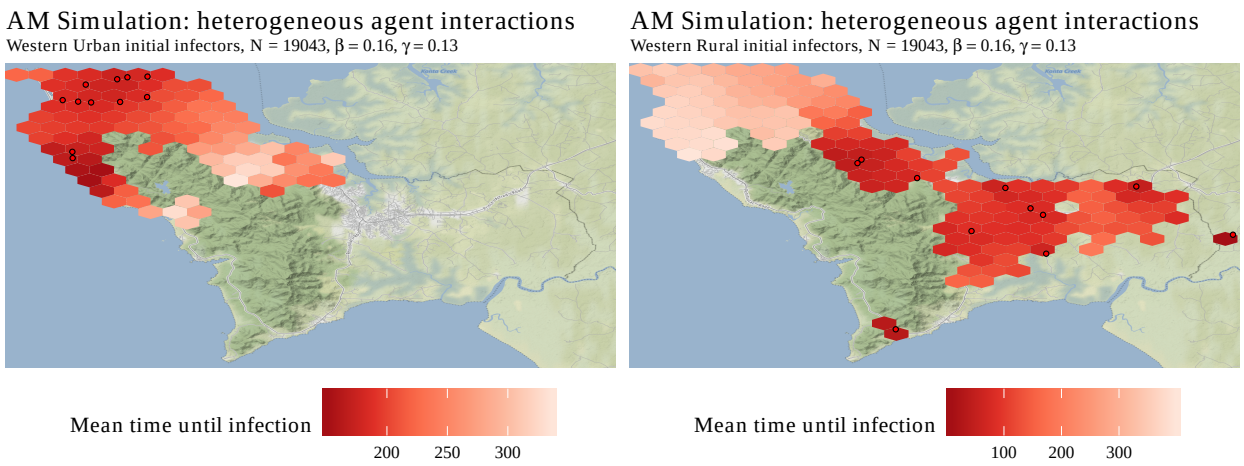

Figure 9.5: Maps of the average time to infection for heterogeneous interaction of agents with initial infections in Western Urban (left) and Western Rural (right).

We plot the average time to infection maps in Figure 9.5. From these maps, we see that the spread of infection differs by the initial infectors location. For the initial infectors within Western Urban, the disease barely spreads into Western Rural. However, since there are many more people in Western Urban, the spread of the disease is more likely to last longer and infect more people over all. For initial infectors located exclusively in Western Rural, we are more likely to see clusters of infections around the initial infectors that very slowly spread outwards. While the disease can spread into Western Urban and travel quite a far distance, the final size of the disease is expected to be only to be 2.94% compared to 21% for Western Urban initial infections.

The differences in our experiment are stark. Initial infections in Western Urban lead to more deadly outbreaks that are centralized and do not travel far over the course of 400 days. In contrast, initial infections in Western Rural mean that the disease will likely die out, but the disease is likely to travel further.

For both sets of initial infections, we see that whenever there are two pairs of agents very close to one another in distance, we have early outbreaks in the close-by regions. Another take away from this experiment is that the 95% CIs for the summary statistics of final size, epidemic duration, peak infection size, and peak day of infection are all much wider than under the assumption of homogeneous interaction of agents, which tells us that the number of contacts/neighbors is also important, along with the effective population size.

9.4 Examining the effective population and contact sizes

In the previous chapter (Ch. 8), we saw that the effective population size N has an important role in shaping our resulting simulations whether that be through direct estimation of disease parameters or the role it has on the variance of the statistics about an outbreak including final size, epidemic duration, peak infectious

size, and day of peak of infectious. In the previous section (Sec. 9.3), we also glimpsed into the role that the maximum number of contacts an agent plays in effecting the results of an Ebola outbreak.

Both the effective population size N and number of maximum neighbors, which we denote as M, influence our CM-AM pair in theoretical and practical ways. As described in Ch. 8, the effective population size may be interpreted as a group of people that interacts approximately homogeneously with infectious individuals. Theoretically, the rate of convergence in probability of the estimate of the number of individuals in each state assuming a deterministic probability of transition (see Eq. (2.4)) is $O_P(\sqrt{N})$ and so large N will result in smaller CIs.

The maximum number of contacts an agent may have also behaves similarly to N and larger M may result in smaller CIs. Besides being the maximum number of contacts we may reasonably assume an agent to have, M also sets limits on the interaction structure of agents. In general, the interaction structure may be very complicated with features such as separate groups, bottle necks, "hub" vertices. In this situation, however, we only know that a single agent can only infect at most $M/N \times 100\%$ of the population.

The practical concerns of M and N are also important because they deal with the computer run time and computer memory required to run simulations. The run time of a simulation is $O\left((T - t_0) \times S(t) \times I(t)\right)$ since our CM-AM pairs are contingent on susceptible-infectious interactions. In the case of Ebola, $S(t) >> I(t)$ but for measles in Hagelloch, $O(S(t)) \approx O(I(t))$. Memory also cannot be neglected as at least $N \times 3$ agent states may be stored in the form of the $\mathbf{U}$ sufficient statistic (see Eq. (5.1)) and more importantly neighbor/contact lists are pre-computed and stored in objects such as a linked list, dictionary, or $N \times M$ array, for instance.

In this set of experiments we

1. Fix β, γ, and L
2. Vary N and M with $N > M$
3. Run the CM-AM pair using the set of initial parameters and heterogeneous agent interaction
4. Report summary statistics and their variation along with time and memory of the neighbor list.

The results of this experiment are displayed in Table 9.1 and show the mean estimates along with the sample error of a number of summary statistics. We see that the mean estimate of final size of the outbreak tends to increase both with increasing N and increasing M. In contrast, the mean estimate of peak infectious seems to be more dependent on the maximum number of neighbors than it does on the effective population size. The mean estimate of epidemic duration is fairly constant once $N > 1000$, but the standard error is always large. The mean estimate of day of peak is also fairly constant and has a smaller standard error than epidemic duration.

Table 9.1: Table of results from simulations with different effective population size N and max number of contacts M.

N	M	M/N	Final Size (%)	SE(Final Size)	Epi. Duration	SE(Epi. Dur.)	Peak I (%)	SE(Peak I)	Day of Peak	SE(Day of Peak)
853	100	0.12	14.23	5.95	282.27	240.24	2.22	0.72	217.58	18.30
853	500	0.59	14.19	5.95	278.93	234.52	2.20	0.70	218.05	18.02
853	852	1.00	14.19	5.95	278.93	234.52	2.20	0.70	218.05	18.02
853	852	1.00	14.19	5.95	278.93	234.52	2.20	0.70	218.05	18.02
8706	100	0.01	18.98	9.42	516.68	316.07	1.09	0.57	335.42	84.91
8706	500	0.06	21.26	9.76	516.55	312.11	1.24	0.57	343.70	86.94
8706	1000	0.11	19.08	10.42	503.63	319.24	1.12	0.59	339.06	91.66
8706	8705	1.00	21.66	9.73	515.16	305.44	1.25	0.59	339.73	69.33
21576	100	0.00	20.35	9.35	552.54	314.16	1.04	0.51	372.99	90.73
21576	500	0.02	22.29	9.72	555.94	311.27	1.10	0.49	372.11	81.84
21576	1000	0.05	24.48	8.96	565.76	290.96	1.27	0.50	377.11	74.96
21576	10000	0.46	23.10	10.42	541.55	314.39	1.27	0.62	351.73	72.00
42373	100	0.00	22.19	9.60	570.63	299.83	1.06	0.47	406.14	84.28
42373	500	0.01	23.35	10.36	564.70	308.85	1.14	0.52	405.71	85.86
42373	1000	0.02	23.48	10.92	561.31	310.25	1.14	0.55	393.52	85.64
42373	10000	0.24	24.71	10.24	573.74	290.68	1.29	0.56	395.06	77.45
63239	100	0.00	19.23	11.69	545.27	329.15	0.98	0.57	424.56	116.56
63239	500	0.01	25.03	9.98	574.14	299.89	1.21	0.50	416.98	79.71
63239	1000	0.02	23.00	11.65	569.77	302.61	1.11	0.57	408.44	85.14
63239	10000	0.16	24.30	11.80	573.12	297.55	1.30	0.64	406.27	81.03
83502	100	0.00	24.01	7.75	584.30	277.07	1.15	0.37	432.82	70.85
83502	500	0.01	23.86	10.17	557.31	319.84	1.21	0.52	397.13	85.42
83502	1000	0.01	25.19	9.78	562.00	317.31	1.27	0.50	406.90	86.10
83502	10000	0.12	27.94	7.15	586.41	281.50	1.50	0.41	406.07	57.56

Overall, we find that the ratio of M to N seems to be more important in estimates than the numbers themselves. Notably we find that 1) the CIs tend to be large and 2) the 95% CIs overlap with one another for every estimate. This seems to indicate that a smaller-scale AM, in terms of N, may be used in lieu of one that attempts to mimic an entire population. Another feature of note is that the sample errors of the estimates do not always decrease as the effective population size increases. The maximum number of contacts seems to have an important role in determining the SEs, as well as proportion of the maximum number of contacts. In future work, it would be interesting to analyze contact structures besides the one based upon uniform selection given the agent is within a one mile radius of the reference agent.

9.4.1 Computer time and memory

In terms of computer time and memory, N and M are very important, and their importance is demonstrated in the results shown in Figure 9.6. Each simulation was run for $L = 100$ runs, and the time reported includes generating the list of pre-computed agent contacts, which is then used in each of the L simulations. In the time vs. $\log N$ graph we find that the required time appears to grow quadratically and the maximum neighbor size is associated with larger run time. The difference between running an AM with 83502 agents for $M = 100$ and $M = 10000$ is close to one hour. In terms of memory, generating pre-computed lists of neighbors is expensive. For the simulation with $N = 83502$ agents and $M = 10^5$, the contact list was 1.5 GB where the median number of contacts each agent had was 4846.

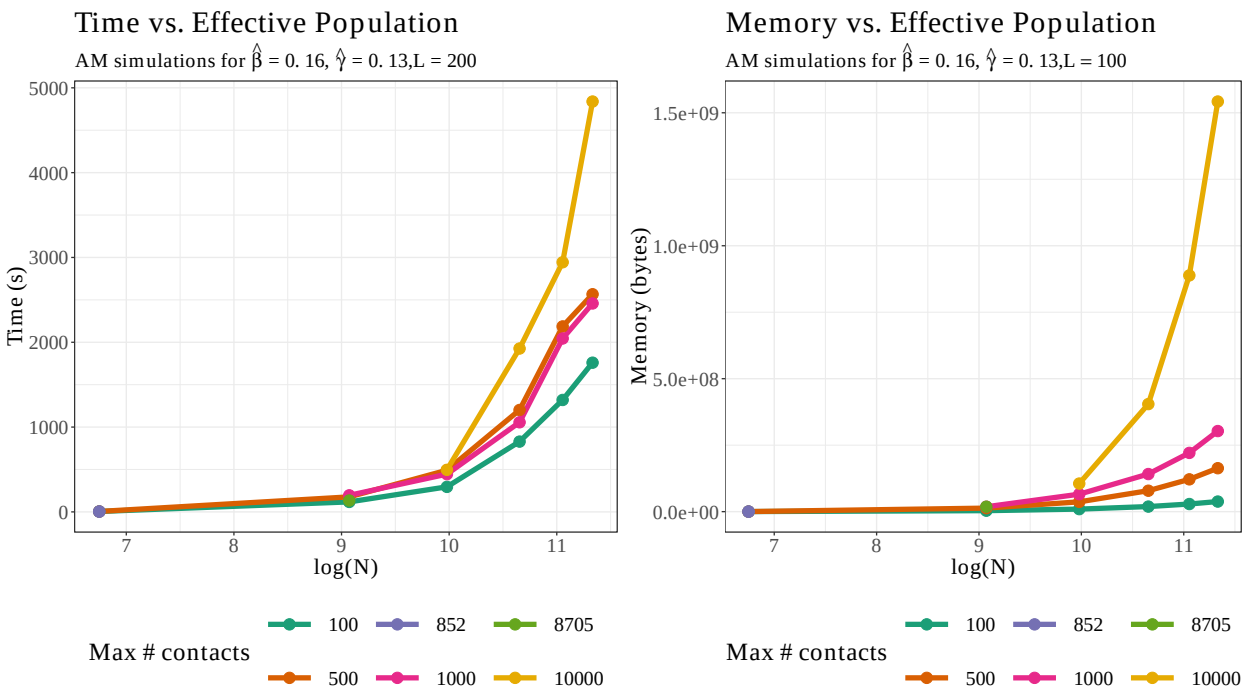

Figure 9.6: Scatter plots of computer time (left) and Memory (right) vs. $\log(N)$ where the lines and points are colored by the maximum number of neighbors used in the simulation.

What this tells us is the following: 1) running a full-scale AM with $N = 1.4$ million would be possible, yet slow, even with the aid of a supercomputer, 2) running a full-scale AM at the country level ($N = 100+$ million) with travel patterns and full agent interaction would be all but impossible, and 3) sub-sampling the population and running the AM on that sub-sample may produce similar and adequate results but is dependent on the assumed interaction structure on the agents.

Basically, modeling a *full scale, full-interaction* CM-AM of a worldwide epidemic is currently an impossible task due to the computer time and memory required to run simulations. Thus for future work, we will explore more ramifications of sub-sampling populations along with using stochastic CM models in conjunction with stochastic AM models to reduce the number of necessary of agent interactions.

9.5 Chapter summary

In this chapter we examine CM-AM pair simulations using the parameters and model selection from Ch. 8 to guide our initial parameter selection. We then examine the effects of varying different initial parameters of the CM-AM pair, specifically, heterogeneous vs. homogeneous agent interaction in terms of both the number of individuals in each state over time as well as the temporal transmission of the disease, initial infection in Western Urban or Western Rural, and the role the effective population size N and the number

of neighbors/contacts we allow in the simulations. Since our data does not lend itself well to all of these situations, many of the results are illustrative of what *could* be done with a CM-AM pair given more information about how the disease is transmitted from one individual to another than opposed to conclusive results about this specific outbreak of Ebola.

That said, we do find our model fitting from the previous chapter to be extremely useful not only in guiding initial parameter estimates but also for determining the number of groups for the disease parameters. Moreover, we find the stochastic CM-AM pair with $N \approx 19000$, $\hat{\beta} = 0.159$, and $\hat{\gamma} = 0.127$, fits the data very well, as shown in Figure 9.1. If researchers were interested in learning about implementing preventions at the level of Western District such as uniformly reducing β or uniformly vaccinating the population of that area, then we recommend using this model.

When we implement a basic interaction scheme based only on physical distance of the agents, we find that the spatial spread of the disease becomes very important. In Figures 9.2 and 9.3 we find that as expected the homogeneous interaction of agents does not depend on the location of agents but once we implement a basic interaction scheme, a very clear pattern emerges such that the infection travels from the northwest part of the district to the southwest part of the district over time. We offer a way to quantify the spatial spread of the disease with the statistic presented in Eq. (9.1), which summarizes the minimum distance between an infected case and the set of initial infectors. When these statistics are used in conjunction with the time until infection, we can compare the relationship between the two, for example, using a Loess smoother as shown in Figure 9.4.

With regards to the initial sensitivity of infection locations, we find that both the summary statistics of the disease (e.g. final size, epidemic duration, peak infectious, day of peak infectious) and the spatial spread of the disease are very different from one another, given our heterogeneous interaction scheme. Although we acknowledge it is unrealistic that people can only spread the disease to those in a one mile radius, this shows that the both contacts of the agents and the location of the agents greatly influence how an epidemic can spread. For example, we find that more population dense areas are more likely to result in a worse outbreak but in some respects are better at containing the disease within the region. On the other hand, agents in more rural and spread-out areas are less likely to transmit the disease, but when they do, the disease can travel quite far. This may be evidence that it is easier to contain a spread of a disease in a population dense area than one that it is not, which may seems counter intuitive. We do note that this simple analysis does not take into account long-distance behaviors such as car and airplane travel, which would allow for farther travel of the disease, and this should be studied in future work.

Finally, we study the theoretical and practical effects the effective population size N and the contact size M have on our AM simulations. We find that although the mean estimate of final size of the epidemic tends to increase as N increases, the CIs of those estimates overlap, suggesting that it is not necessary to have a

full-scale simulation of a population. Determining, however, what number of agents is needed to have an adequate AM remains an open problem.

Chapter 10

Conclusion and future directions

10.1 Dissertation summary

In this dissertation we address the problem of statistical inference in infectious disease modeling. Specifically, we examine the model classes of compartment models (CM) and agent-based models (AM), and we:

1. Relate statistical properties of CMs and AMs
2. Improve methodology for model selection within the SI and SIR-frameworks
3. Apply our improved theory and methodology to two case studies.

In Chapter 1, we introduce the problem and related work. We introduce epidemic modeling, using the CMs and AMs as model classes. We briefly detail the history of these two classes and how the two classes are sometimes combined together as hybrid models. There are few studies of how the statistical properties of the two models are related to one another, and we improve upon this gap. Following that, we examine related work with respect to parameter estimation and model selection within the CM and AM framework. These include likelihood maximization methods, comparing models through measures of agent interaction, and diagnostic plots. Finally, we examine studies relating to measles and Ebola, which are examined in depth in the dissertation.

In Chapters 2-3, we address the first issue of relating statistical properties of CMs and AMs. We begin by introducing the Kermack and McKendrick (1927) deterministic SIR equations. We then introduce a stochastic CM that incorporates the deterministic SIR equations in a Binomial random draw where the scaled deterministic SIR equations act as probability of transition from one state to the next. In Theorems 2.1-2.2 we calculate the expected value and variance, respectively, of the number of individuals in each state of our Binomial model. In Eq. (2.5), we introduce a stochastic AM with a Bernoulli probability of transition.

In contrast to the stochastic CM, the stochastic AM allows us to track individuals over time whereas the CM only examines the total number of individuals in each state over time. In Theorem 2.3 we state our main theoretical contribution that the Binomial CM and Bernoulli AM are equivalent in distribution, with respect to the number of individuals in each state for each time step.

In Section 2.3, we extend our theory a more general case where we have K total states and the probability of transition matrix $D(t)$ has a specific form. We present a CM and AM with K total states such that the CM and AM are equivalent in distribution in terms of the number of individuals in each state at a given time. Together, we call the equivalent models the CM-AM pair.

In Chapter 3, we examine what are essential features of CMs and AMs to create more general CM-AM pairs. We define a CM to be characterized by homogeneity of individuals within states and homogeneous interaction between individuals in ssuceptible and infectious states. In Section 3.2, we show that any CM has an equivalent AM and explore the consequences of violating the assumption of homogeneity of individuals within groups. We also show in an example we call the "lock-step" model that independence of individuals is not a requirement to have an equivalent AM. We then show that every AM has an equivalent CM, provided we adjust the total number of states. The equivalence between general CMs and AMs demonstrates the importance of the total number of states to use in an epidemic model and hence is an important step in model selection.

We address the second issue of improvement in parameter and model selection in Chapter 4. We detail novel methods to aid in selection of the total number of states to use in an epidemic CM or AM provided the disease-level states are given. Specifically, for models in the SIR-framework, we present two novel visual diagnostics to use in conjunction with other model selection techniques in order to select the best model. The plots include transforming SIR data so that the slope of the best fit line corresponds to $\mathcal{R}_0$, the reproduction number. This along with weighted linear regression through a plug-in variance estimate yields an empirical 95% coverage in our transformed SIR model. The second plot presents SIR data in the format of a ternary plot that includes confidence regions and time scales to assess the fit of the data. Additionally, for the SI-framework, we provide a statistical investigation specific to models within the SI-framework to quantify whether individuals interact homogeneously "enough" in the sense that we can use a CM with fewer total states in contrast to a CM or AM with more states and can still adequately model the epidemic.

In Chapters 5-9 we apply our theory and methodology to two real-world scenarios. The first scenario is an outbreak of measles in Hagelloch Germany in the 1860s, and the second scenario is the Ebola outbreak of 2014-2015 in Western District, Sierra Leone. The two outbreaks differ in scale (the Hagelloch outbreak is much smaller than the Western District outbreak) and data features (the Hagelloch data has more detailed demographic and interaction features than the Western District outbreak), and we analyze our methodology in view of these two extremes.

In Chapter 5, we explore the Hagelloch data, especially in regard to individual interaction features such as household location and school class. We then fit models to the data and find that there is a difference in infection rates before and day after 25 of the outbreak. The model we select has six total states and two groups of individuals and estimate that the reproduction number for the epidemic 4.94 (95% CI: [4.68, 5.21]), assuming all individuals act interact homogeneously and have the same rates of recovery and infection. This estimate of $\mathcal{R}_0$ is much lower than the estimates presented in Anderson and May (1992) but is comparable to the estimates of $\mathcal{R}_0$ in Getz et al. (2016).

Following our model selection and parameter estimation for the Hagelloch measles outbreak, in Chapters 6-7, we explore scenarios of our model using our CM-AM pair that may be examined when planning prevention and intervention routines. In Chapter 6, we examine the more abstract scenario of the effects of reducing β_k, the infectivity of the disease for group k, without delving into the specifics of how β_k is reduced. We find that we can better prevent large-scale outbreaks by implementing reductions in β_k sooner, rather than later. More specifically, we find that we need to reduce β_k by at least 50% to obtain a 25% reduction in the final size of the epidemic. In our simulations, we also show that it is worthwhile in terms of final size of the epidemic to implement preventions even one month into the epidemic.

In Chapter 7, we look at more tangible prevention routines with our AM which include individual isolation and quarantine along with school closure. Again we find that we need to reduce β_k by around 50% to see a significant reduction in the final size of the epidemic (20% reduction in final size). We find isolation and quarantine to be particularly effective as intervention techniques, even when allowing for a short delay between the infectious period and beginning of isolation. One problem we repeatedly find is that the CIs for our estimates of peak percent infectious, day of peak percent infectious, epidemic duration, and final size have very large CIs, sometimes spanning the whole space of the estimate. One reason for this is small population size and another important reason is that the probability of transition in the model we are using is itself a random variable.

In Chapters 8-9, we examine the Ebola outbreak in Western District, Sierra Leone. The Western District data is feature-poor compared to the Hagelloch data because we only have infection dates and age of infected cases whereas in the Hagelloch data we have household information, school class information, and even the purported infector ID. As such, we pursue more high level concepts as opposed to answering specific questions about the Ebola outbreak. Specifically, we examine the importance of the value of the effective population size N, sensitivity of the model results to initial locations of the first infections, and individual interaction restrictions.

In Chapter 8 we examine the data, perform model selection, and estimate disease parameters. In contrast to the Hagelloch outbreak, we also examine the importance of the effective population size N has on our model. We select our best model to have one group of individuals with $\beta = 0.16$, $\gamma = 0.12$ and $N =$

18, 768, which in context means that there is a community of approximately 18,768 individuals that behave approximately homogeneously with infected agents.

In Chapter 9, we use our best fit CM-AM pair to examine the scenarios of homogeneous versus heterogeneous agent interactions, sensitivity to initial location of infectors, and effective population size along with the number of contacts each agent is assumed to have. We show how our AM can include spatial spread and how important the transmission of the disease between individuals in our model. We also show that the effective population size and number of contacts each individual has is very important both in terms of modeling results but also in practice in terms of computer time and memory required.

Finally, in Chapter 10 we summarize the results of our dissertation and provide directions for future work.

10.2 Future directions

Future directions include all three issues addressed in this dissertation: theoretical, methodological, and practical.

Theoeretical questions include:

- Can we examine specific non-Binomial or Multinomial CMs? How does this effect the equivalent AM?
- Exploring the bias-variance tradeoff when selecting K^*, the optimal number of states used to model the epidemic
- Using priors for the probability of transition, beginning with a conjugate β and extending to non-parametric priors
- More general expected value and variance calculations for CMs where the probability of transition follows either typical ODEs or Reed-Frost transitions
- How can we include the effective population size as a random variable in our models?

Methodological questions include:

- Can we extend the concept of quantifying whether populations are homogeneous enough to frameworks beyond the SI disease-level states?
- Can we extend the ternary plot to a 3D plot for use with the SEIR, which is a very common choice in disease modeling?

Pratical questions include:

- Model a present day measles outbreak using a heterogeneous population

- Add hospitals and ETUs to the Ebola modeling
- Examine ring trials where we vaccinate contacts of the infectious and contacts of contacts

There are still many aspects to explore and refine when it comes to epidemic modeling with CM or AM pairs, and we hope that the work presented here can be used to help prevent and eradicate infectious disease.

Bibliography

Abar, S., Theodoropoulos, G. K., Lemarinier, P., and O'Hare, G. M. (2017). Agent based modelling and simulation tools: A review of the state-of-art software. *Computer Science Review*, 24:13 – 33. 6

Abbey, H. (1952). An examination of the reed-frost theory of epidemics. *Human Biology*, 24(3):201. 3, 11, 15, 18, 59, 74

Adamatzky, A. (2010). *Game of Life Cellular Automata*. Springer Publishing Company, Incorporated, 1st edition. 4

Ajelli, M., Merler, S., Fumanelli, L., Pastore y Piontti, A., Dean, N. E., Longini, I. M., Halloran, M. E., and Vespignani, A. (2016). Spatiotemporal dynamics of the ebola epidemic in guinea and implications for vaccination and disease elimination: a computational modeling analysis. *BMC Medicine*, 14(1):130. 11

Allen, L. J. (1994). Some discrete-time si, sir, and sis epidemic models. *Mathematical Biosciences*, 124(1):83 – 105. 121

Allen, L. J. and Burgin, A. M. (2000). Comparison of deterministic and stochastic {SIS} and {SIR} models in discrete time. *Mathematical Biosciences*, 163(1):1 – 33. 3

Althaus, C. L. (2014). Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. *PLOS Current Outbreaks*. 3

Althaus, C. L. (2015). Rapid drop in the reproduction number during the ebola outbreak in the democratic republic of congo. *PeerJ*, 3:e1418. 11

Anderson, R. and May, R. (1992). *Infectious Diseases of Humans*. Oxford: Oxford University Press. 3, 11, 15, 18, 50, 69, 78, 121, 141

Anderson, R., Medley, G., May, R., and Johnson, A. (1986). A preliminary study of the transmission dynamics of the human immunodeficiency virus (hiv), the causative agent of aids. *Mathematical Medicine and Biology*, 3(4):229–263. 4

Axtell, R., Axelrod, R., Epstein, J. M., and Cohen, M. D. (1996). Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory*, 1(2):123–141. 7, 15

Backer, J. A. and Wallinga, J. (2016). Spatiotemporal Analysis of the 2014 Ebola Epidemic in West Africa. *PLOS Computational Biology*, 12(12):1–17. 12, 116, 123, 124

Baize, S., Pannetier, D., Oestereich, L., Rieger, T., Koivogui, L., Magassouba, N., Soropogui, B., Sow, M. S., Keïta, S., De Clerck, H., Tiffany, A., Dominguez, G., Loua, M., Traoré, A., Kolié, M., Malano, E. R., Heleze, E., Bocquin, A., Mély, S., Raoul, H., Caro, V., Cadar, D., Gabriel, M., Pahlmann, M., Tappe, D., Schmidt-Chanasit, J., Impouma, B., Diallo, A. K., Formenty, P., Van Herp, M., and Günther, S. (2014). Emergence of Zaire Ebola Virus Disease in Guinea. *New England Journal of Medicine*, 371(15):1418–1425. PMID: 24738640. 114

Bajardi, P., Poletto, C., Ramasco, J. J., Tizzoni, M., Colizza, V., and Vespignani, A. (2011). Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PLOS ONE*, 6(1):1–8. 5, 8

Balser, E. (2019). Measles confirmed in Pittsburgh, health officials say. https://triblive.com/local/pittsburgh-allegheny/measles-confirmed-in-pittsburgh-health-officials-say/. 68

Banos, A., Corson, N., Gaudou, B., Laperrière, V., and Coyrehourcq, S. R. (2015). The importance of being hybrid for spatial epidemic models: A multi-scale approach. *Systems*, 3(4):309–329. 3, 8

Barrett, C., Bisset, K., Chandan, S., Chen, J., Chungbaek, Y., Eubank, S., Evrenosoğlu, Y., Lewis, B., Lum, K., Marathe, A., et al. (2013). Planning and response in the aftermath of a large crisis: An agent-based informatics framework. In *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, pages 1515–1526. IEEE Press. 5

Barrett, C. L., Bisset, K. R., Eubank, S. G., Feng, X., and Marathe, M. V. (2008). Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, page 37. IEEE Press. 6

Becker, A. D. and Grenfell, B. T. (2017). tsiR: An R package for time-series Susceptible-Infected-Recovered models of epidemics. *PLOS ONE*, 12(9):1–10. 16

Becker, N. (1981). A general chain binomial model for infectious diseases. *Biometrics*, 37(2):251–258. 3

Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., and Dhiman, R. C. (2011). Malaria in Northwest India: Data Analysis via Partially Observed Stochastic Differential Equation Models Driven by Lévy Noise. *Journal of the American Statistical Association*, 106(494):440–451. 11

Bobashev, G. V., Goedecke, D. M., Yu, F., and Epstein, J. M. (2007). A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches. In *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, pages 1532–1537. IEEE Press. 7

Bradhurst, R. A., Roche, S. E., East, I. J., Kwan, P., and Garner, M. G. (2015). A hybrid modeling approach to simulating foot-and-mouth disease outbreaks in Australian livestock. *Frontiers in Environmental Science*, 3:17. 8

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., and Rosenfeld, R. (2015). Flexible modeling of epidemics with an empirical bayes framework. *PLOS Computational Biology*, 11(8):1–18. 8, 9, 92

Brown, G. D., Oleson, J. J., and Porter, A. T. (2016). An empirically adjusted approach to reproductive number estimation for stochastic compartmental models: A case study of two ebola outbreaks. *Biometrics*, 72(2):335–343. 11

Capaldi, A., Behrend, S., Berman, B., Smith, J., Wright, J., and Lloyd, A. L. (2012). Parameter estimation and uncertainty quantication for an epidemic model. *Mathematical Biosciences and Engineering*, page 553. 92

Carley, K. M., Fridsma, D. B., Casman, E., Yahja, A., Altman, N., Chen, L.-C., Kaminsky, B., and Nave, D. (2006). Biowar: scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 36(2):252–265. 15

Centers for Disease Control and Prevention (2018). Measles history. Available online at `https://www.cdc.gov/measles/about`. 68, 69

Centers for Disease Control and Prevention (2019). Ebola virus disease. Available online at `https://www.cdc.gov/vhf/ebola/transmission/index.html`. 115, 117

Chao, D. L., Halloran, M. E., Obenchain, V. J., and Longini, Jr, I. M. (2010). Flute, a publicly available stochastic influenza epidemic simulation model. *PLOS Computational Biology*, 6(1):1–8. 6, 9

Chen, L.-C., Kaminsky, B., Tummino, T., Carley, K. M., Casman, E., Fridsma, D., and Yahja, A. (2004). *Aligning Simulation Models of Smallpox Outbreaks*, pages 1–16. Springer Berlin Heidelberg, Berlin, Heidelberg. 7

Chen, T. M. and Jamil, N. (2006). Effectiveness of quarantine in worm epidemics. In *2006 IEEE International Conference on Communications*, volume 5, pages 2142–2147. IEEE. 92

Chris, G., David, W., and R., H. D. (2012). A network-based analysis of the 1861 hagelloch measles data. *Biometrics*, 68(3):755–765. 11

Cohen, J. (2019). Ebola outbreak continues despite powerful vaccine. *Science*, 364(6437):223–223. 114

Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103(7):2015–2020. 4, 9, 49

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3):553–570. 6, 8, 9

Daley, D. J., Gani, J., and Gani, J. M. (2001). *Epidemic modelling: an introduction*, volume 15. Cambridge University Press. 3, 18, 31, 32, 59

Drake, J. M., Bakach, I., Just, M. R., O'Regan, S. M., Gambhir, M., and Fung, I. C.-H. (2015). Transmission models of historical ebola outbreaks. *Emerging infectious diseases*, 21(8):1447. 121

Drake, N. (2019). The world's second-biggest Ebola outbreak is still raging. Here's why. https://www.nationalgeographic.com/science/2019/04/worlds-second-biggest-ebola-outbreak-still-raging-heres-why-hot-zone/. 114

Edwards, M., Huet, S., Goreaud, F., and Deffuant, G. (2003). Comparing an individual-based model of behaviour diffusion with its mean field aggregate approximation. *Journal of Artificial Societies and Social Simulation*, 6(4). 7

Epstein, J. M. (2007). Agent-based computational models and generative social science [generative social science studies in agent-based computational modeling]. *Introductory Chapters*. 2, 4, 91, 92

Eubank, S., Barrett, C., Beckman, R., Bisset, K., Durbeck, L., Kuhlman, C., Lewis, B., Marathe, A., Marathe, M., and Stretz, P. (2010). Detail in network models of epidemiology: are we there yet? *Journal of biological dynamics*, 4(5):446–455. 5, 28

Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184. 5, 15

Fahse, L., Wissel, C., and Grimm, V. (1998). Reconciling classical and individual-based approaches in theoretical population ecology: A protocol for extracting population parameters from individual-based models. *The American Naturalist*, 152(6):838–852. PMID: 18811431. 7

Figueredo, G. P., Siebers, P.-O., Owen, M. R., Reps, J., and Aickelin, U. (2014). Comparing stochastic differential equations and agent-based modelling and simulation for early-stage cancer. *PLOS ONE*, 9(4):1–18. 3, 7

Fintzi, J., Cui, X., Wakefield, J., and Minin, V. N. (2017). Efficient data augmentation for fitting stochastic epidemic models to prevalence data. *Journal of Computational and Graphical Statistics*, 0(ja):0–0. 3

Gallagher, S., Chang, A., and Eddy, W. F. (2019). Nine ways to estimate R_0 in the SIR model with applications to 2009 pandemic influenz. *In preparation*. 9, 90

Gallagher, S., Richardson, L. F., Ventura, S. L., and Eddy, W. F. (2018). SPEW: Synthetic Populations and Ecosystems of the World. *Journal of Computational and Graphical Statistics*. 6, 120

Gani, J. and Yakowitz, S. (1995). Error bounds for deterministic approximations to markov processes, with applications to epidemic models. *Journal of Applied Probability*, 32(4):1063–1076. 3

Getz, W. M., Carlson, C., Dougherty, E., Porco, T. C., and Salter, R. (2016). An agent-based model of school closing in under-vacccinated communities during measles outbreaks. In *Proceedings of the Agent-Directed Simulation Symposium*, page 10. Society for Computer Simulation International. 11, 69, 78, 90, 141

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403 – 434. 3

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., Wheaton, W. D., Fyshe, A., Galloway, D. D., Sriram, A., Guclu, H., Abraham, T., and Burke, D. S. (2013). 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):1–14. 5, 15, 92, 109

Groendyke, C. and Welch, D. (2018). epinet: An r package to analyze epidemics spread across contact networks. *Journal of Statistical Software, Articles*, 83(11):1–22. 68

Hanski, I. (1998). Metapopulation dynamics. *Nature*, 396(6706):41–49. 8

Harko, T., Lobo, F. S., and Mak, M. (2014). Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation*, 236:184–194. 10, 50, 64

He, D., Ionides, E. L., and King, A. A. (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of The Royal Society Interface*. 9, 11, 18

Health and Human Services (2019). What is the difference between isolation and quarantine? https://www.hhs.gov/answers/public-health-and-safety/what-is-the-difference-between-isolation-and-quarantine/index.html. 102

Henao-Restrepo, A. M., Camacho, A., Longini, I. M., Watson, C. H., Edmunds, W. J., Egger, M., Carroll, M. W., Dean, N. E., Diatta, I., Doumbia, M., Draguez, B., Duraffour, S., Enwere, G., Grais, R., Gunther, S., Gsell, P.-S., Hossmann, S., Watle, S. V., Kondé, M. K., Kéïta, S., Kone, S., Kuisma, E., Levine, M. M., Mandal, S., Mauget, T., Norheim, G., Riveros, X., Soumah, A., Trelle, S., Vicari, A. S., Røttingen, J.-A., and Kieny, M.-P. (2017). Efficacy and effectiveness of an rvsv-vectored vaccine in preventing ebola virus disease: final results from the guinea ring vaccination, open-label, cluster-randomised trial (ebola Ça suffit!). *The Lancet*, 389(10068):505 – 518. 12, 92, 115

Hethcote, H. W. (1994). A thousand and one epidemic models. In Levin, S. A., editor, *Frontiers in Mathematical Biology*, pages 504–515, Berlin, Heidelberg. Springer Berlin Heidelberg. 4

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653. 3

Hooten, M. B., Anderson, J., and Waller, L. A. (2010). Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology*, 1(2):177 – 185. GEOMED Conference. 4

Hooten, M. B. and Wikle, C. K. (2010). Statistical agent-based models for discrete spatio-temporal systems. *Journal of the American Statistical Association*, 105(489):236–248. 6

Hunter, E., Mac Namee, B., and Kelleher, J. (2018). An open-data-driven agent-based model to simulate infectious disease outbreaks. *PLOS ONE*, 13(12):1–35. 9, 11

Jacquez, J. A. and O'Neill, P. (1991). Reproduction numbers and thresholds in stochastic epidemic models i. homogeneous populations. *Mathematical Biosciences*, 107(2):161 – 186. 3

Jaffry, S. W. and Treur, J. (2008). Agent-based and population-based simulation: A comparative case study for epidemics. In *Proceedings of the 22nd European Conference on Modelling and Simulation*, pages 123–130. Citeseer. 8

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721. 2, 3, 16, 31, 50, 139

King, A. A., Nguyen, D., and Ionides, E. L. (2015). Statistical inference for partially observed markov processes via the r package pomp. *arXiv preprint arXiv:1509.00503*. 9, 16, 32

Koide, C. and Seno, H. (1996). Sex ratio features of two-group SIR model for asymmetrie transmission of heterosexual disease. *Mathematical and Computer Modelling*, 23(4):67 – 91. 31

Lash, T. L. and Fink, A. K. (2003). Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*, 14(4):451–458. 92

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177. 3

Lessler, J., Salje, H., Grabowski, M. K., and Cummings, D. A. T. (2016). Measuring spatial dependence for infectious disease epidemiology. *PLOS ONE*, 11(5):1–13. 68

Leventhal, G. E., Günthard, H. F., Bonhoeffer, S., and Stadler, T. (2013). Using an Epidemiological Model for Phylogenetic Inference Reveals Density Dependence in HIV Transmission. *Molecular Biology and Evolution*, 31(1):6–17. 58

Lima, A., De Domenico, M., Pejovic, V., and Musolesi, M. (2015). Disease containment strategies based on mobility and information dissemination. *Scientific reports*, 5:10650. 92

Liu, F., Enanoria, W. T., Zipprich, J., Blumberg, S., Harriman, K., Ackley, S. F., Wheaton, W. D., Allpress, J. L., and Porco, T. C. (2015a). The role of vaccination coverage, individual behaviors, and the public health response in the control of measles epidemics: an agent-based simulation for california. *BMC Public Health*, 15(1). 5, 6, 11, 68

Liu, S., Pang, L., Ruan, S., and Zhang, X. (2015b). Global dynamics of avian influenza epidemic models with psychological effect. *Computational and Mathematical Methods in Medicine*, 2015. 15

Longini, Jr., I. M., Halloran, M. E., Nizam, A., and Yang, Y. (2004). Containing pandemic influenza with antiviral agents. *American Journal of Epidemiology*, 159(7):623. 5

Mills, C. E., Robins, J. M., and Lipsitch, M. (2004). Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906. 3

Mpeshe, S. C., Nyerere, N., and Sanga, S. (2017). Modeling approach to investigate the dynamics of zika virus fever: A neglected disease in africa. *Int. J. Adv. Appl. Math. and Mech*, 4(3):14–21. 50

Nakamura, G. M., Monteiro, A. C. P., Cardoso, G. C., and Martinez, A. S. (2017). Efficient method for comprehensive computation of agent-level epidemic dissemination in networks. *Scientific reports*, 7:40885. 8, 11

Neal, P. J. and Roberts, G. O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, 5(2):249–261. 10, 69, 70, 157

Nishiura, H. and Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In Chowell, G., Hyman, J. M., Bettencourt, L. M. A., and Castillo-Chavez, C., editors, *Mathematical and Statistical Estimation Approaches in Epidemiology*, pages 103–121. Springer Netherlands, Dordrecht. 92

Oesterle, H. (1992). Statistische Reanalyse einer Masernepidemie 1861 in Hagelloch. 68

Pandey, A., Atkins, K. E., Medlock, J., Wenzel, N., Townsend, J. P., Childs, J. E., Nyenswah, T. G., Ndeffo-Mbah, M. L., and Galvani, A. P. (2014). Strategies for containing Ebola in West Africa. *Science*, 346(6212):991–995. 3, 11

Pfeilsticker, A. (1863). Beiträge zur Pathologie der Masern mit besonderer Berücksichtigung der statistischen Verhältnisse. 68

Rahmandad, H. and Sterman, J. (2008). Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management Science*, 54(5):998–1014. Copyright - Copyright Institute for Operations Research and the Management Sciences May 2008; Document feature - Equations; Graphs; Tables; ; Last updated - 2016-04-02; CODEN - MNSCDI. 7, 9

Rizkalla, C., Blanco-Silva, F., and Gruver, S. (2007). Modeling the Impact of Ebola and Bushmeat Hunting on Western Lowland Gorillas. *EcoHealth*, 4(2):151–155. 11, 50

Rvachev, L. A. and Longini, I. M. (1985). A mathematical model for the global spread of influenza. *Mathematical Biosciences*, 75(1):3 – 22. 4

Safan, M., Heesterbeek, H., and Dietz, K. (2006). The minimum effort required to eradicate infections in models with backward bifurcation. *Journal of mathematical biology*, 53(4):703–718. 10, 55

Salmon, M., Schumacher, D., and Höhle, M. (2016). Monitoring count time series in R: Aberration detection in public health surveillance. *Journal of Statistical Software*, 70(10):1–35. xiii, 69, 70

Scheffer, M., Baveco, J., DeAngelis, D., Rose, K., and van Nes, E. (1995). Super-individuals a simple solution for modelling large populations on an individual basis. *Ecological Modelling*, 80(2):161 – 170. 6

Schelling, T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1. 4

Shrestha, S., King, A. A., and Rohani, P. (2011). Statistical inference for multi-pathogen systems. *PLOS Computational Biology*, 7(8):1–14. 9, 12

Siettos, C., Anastassopoulou, C., Russo, L., Grigoras, C., and Mylonakis, E. (2015). Modeling the 2014 Ebola virus epidemic–agent-based simulations, temporal analysis and future predictions for Liberia and Sierra Leone. *PLoS currents*, 7. 6, 11

Smith, L., Beckman, R., and Baggerly, K. (1995). *TRANSIMS: Transportation analysis and simulation system.* 5

Smith, M. C. and Broniatowski, D. A. (2016). Modeling influenza by modulating flu awareness. In Xu, K. S., Reitter, D., Lee, D., and Osgood, N., editors, *Social, Cultural, and Behavioral Modeling*, pages 262–271, Cham. Springer International Publishing. 50

SPEW (2017). Sierra Leone Synthetic Ecosystem. Available at `http://data.olympus.psc.edu/syneco/spew_1.3.0/africa/western_africa/sle/diags/`. 117

Stobbe, M. (2019). Measles Count in US This Year Already More Than All of 2018. https://www.usnews.com/news/health-news/articles/2019-04-01/measles-count-in-us-this-year-already-more-than-all-of-2018. 10, 68

Venkatramanan, S., Lewis, B., Chen, J., Higdon, D., Vullikanti, A., and Marathe, M. (2018). Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, 22:43 – 49. The RAPIDD Ebola Forecasting Challenge. 9

Vincenot, C. E., Giannino, F., Rietkerk, M., Moriya, K., and Mazzoleni, S. (2011). Theoretical considerations on the combined use of system dynamics and individual-based modeling in ecology. *Ecological Modelling*, 222(1):210 – 218. 7

Wallentin, G. and Neuwirth, C. (2017). Dynamic hybrid modelling: Switching between {AB} and {SD} designs of a predator-prey model. *Ecological Modelling*, 345:165 – 175. 6, 7

Wang, Z., Bauch, C. T., Bhattacharyya, S., d'Onofrio, A., Manfredi, P., Perc, M., Perra, N., Salathé, M., and Zhao, D. (2016). Statistical physics of vaccination. *Physics Reports*, 664:1–113. 5

Waraich, R. A., Charypar, D., Balmer, M., Axhausen, K. W., Waraich, R. A., Waraich, R. A., Axhausen, K. W., and Axhausen, K. W. (2009). Performance improvements for large scale traffic simulation in matsim. In *9th Swiss Transport Research Conference, Ascona*. Citeseer. 5

Wasserman, L., editor (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer. 9, 73

Wheeler, D. C. and Waller, L. A. (2008). Mountains, valleys, and rivers: the transmission of raccoon rabies over a heterogeneous landscape. *Journal of agricultural, biological, and environmental statistics*, 13(4):388. 9

Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media, Inc. 5

Wolkewitz, M., Dettenkofer, M., Bertz, H., Schumacher, M., and Huebner, J. (2008). Statistical epidemic modeling with hospital outbreak data. *Statistics in Medicine*, 27(30):6522–6531. 7

World Heritage Cites (2018). Western Area Peninsula National Park. Available at `http://whc.unesco.org/en/tentativelists/5741/`. 117

Yang, H., Tang, M., and Gross, T. (2015). Large epidemic thresholds emerge in heterogeneous networks of heterogeneous nodes. *Scientific reports*, 5:13122. 7

Zaman, G., Kang, Y. H., and Jung, I. H. (2009). Optimal treatment of an SIR epidemic model with time delay. *Biosystems*, 98(1):43 – 50. 121

Zhao, L., Cui, H., Qiu, X., Wang, X., and Wang, J. (2013). SIR rumor spreading model in the new media age. *Physica A: Statistical Mechanics and its Applications*, 392(4):995 – 1003. 50

Zhou, L., Wang, Y., Xiao, Y., and Li, M. Y. (2019). Global dynamics of a discrete age-structured SIR epidemic model with applications to measles vaccination strategies. *Mathematical Biosciences*, 308:27 – 37. 4, 11

Appendix

Appendix A

Hagelloch EDA

We investigate whether we suspect where and how any heterogeneities in susceptibility or interacting may occur. Neal and Roberts (2004) use heterogeneties in infection based household, class level, and physical distance between pairs of individuals along with a global rate of infection. As such, we focus on those attributes here.

The initial reproduction number $\mathcal{R}_0$ measures the intensity of an initial outbreak by counting the generations of a disease (as opposed to infections over time like in Fig. 5.2). In Figure A.1, we measure the average number of infections generated by the children who become infectious at time t. We see that at about 2 weeks in, we observe a large spike in the average generations produced by a single infectious child.

In Figure A.2, we plot the state of each child over time where the color represents the state of the child at that time. There are two different sets of bars, the first with darker hues and the second with lighter hues. These different sets of bars represent a change in household ID of the children, i.e. children are grouped by household. This graph makes clear that household is a very important variable in modelling the spread of measles since we see that the red sections for each family overlap for nearly all children in all families, meaning that an infection passes through a household in a consecutive amount of time. There are exceptions to this, however. For example, the top most family has 4 children, two of which are infectious at the same time, then a small gap, and then the other two children become infected. This shows that household can account for most of the spread of the disease of children in the same family but not for all of it.

Besides sibling-sibling transmission, within-class transmission is important in the spread of this epidemic. The network of infections is plotted in Figure A.3 where the nodes are children (with location of the nodes overlaid on household location) and the nodes are colored by class. In the bottom of the figure, the network is faceted to show the within-class transmission of the disease. We see that a boy in 1st class purportedly infected 26 of his classmates between November 21 and November 25. For the 1st class, 27 out of 30 were infected by another 1st classmate. For the second class, 48 out of 68 were infected by a direct classmate.

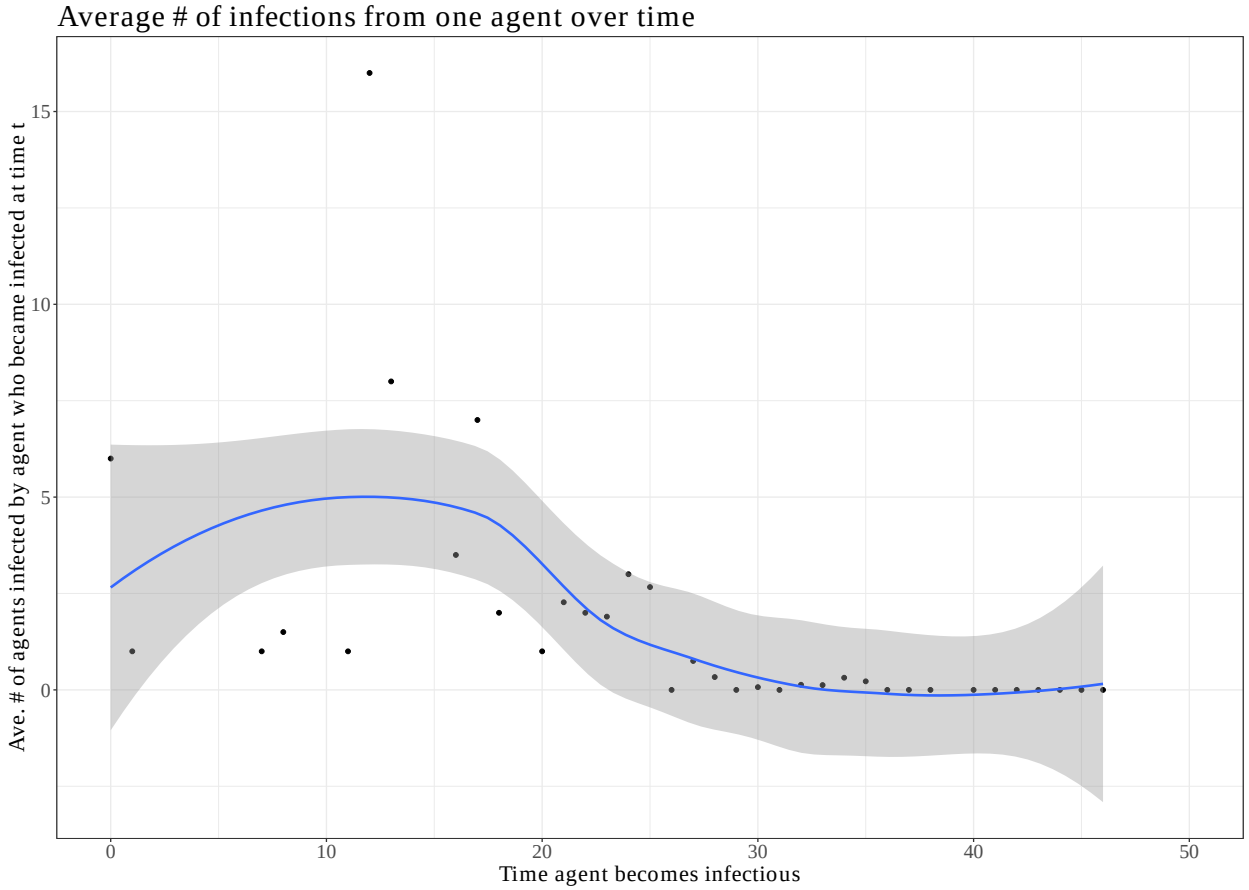

Figure A.1: The average number of infections generated by children who become infectious at time t with a Loess smoother and 95% CI plotted.

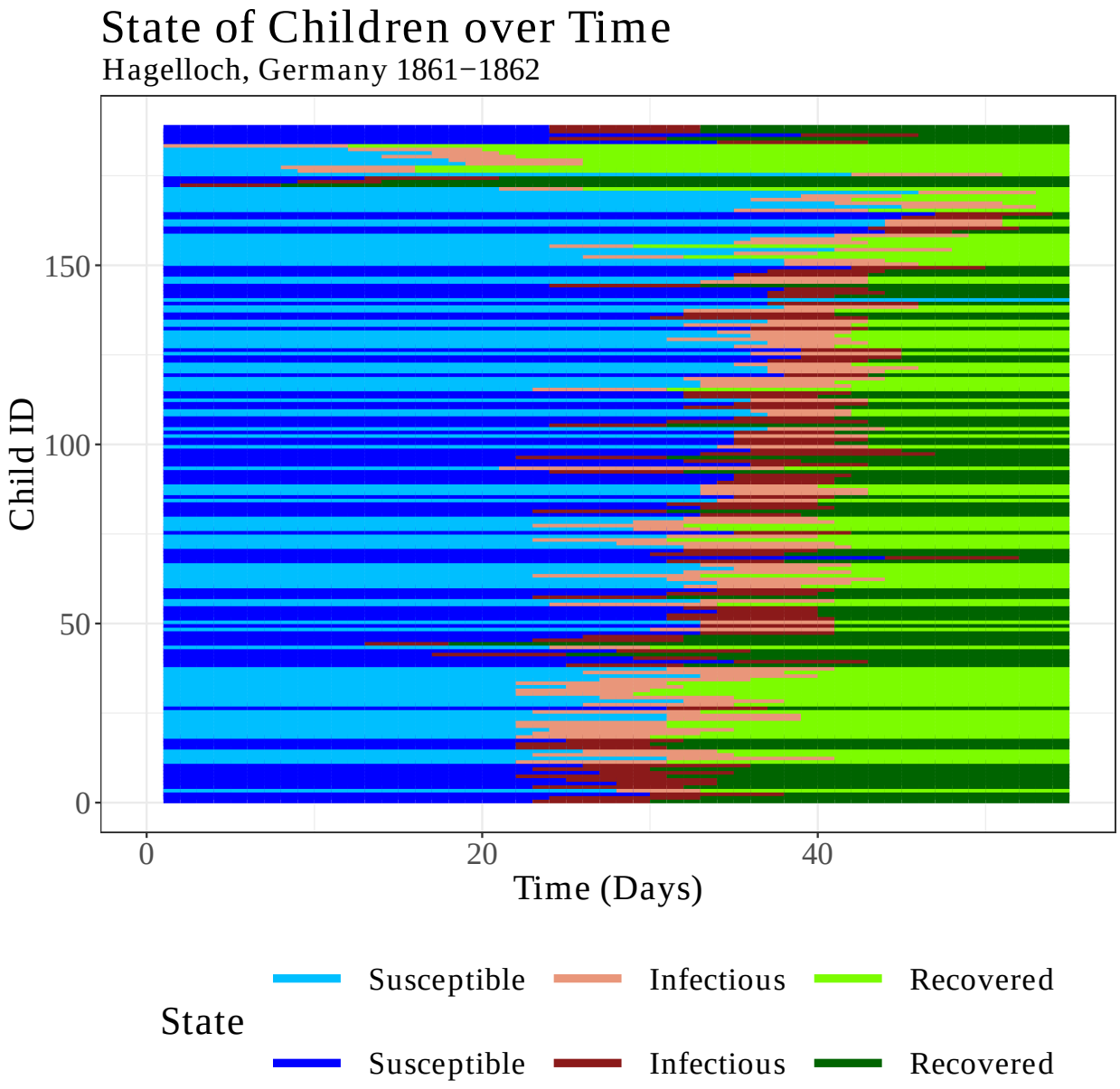

Figure A.2: Each infected child's state is plotted over time. Blue is susceptible, red is infectious, and green is recovered. The alternating shades indicate a change in the household ID of the children.

For the pre-schoolers, only 34 out of 90 children were infected by another pre-schooler. We emphasize that pre-school here means those too young for school and is not a "class" in the usual sense.

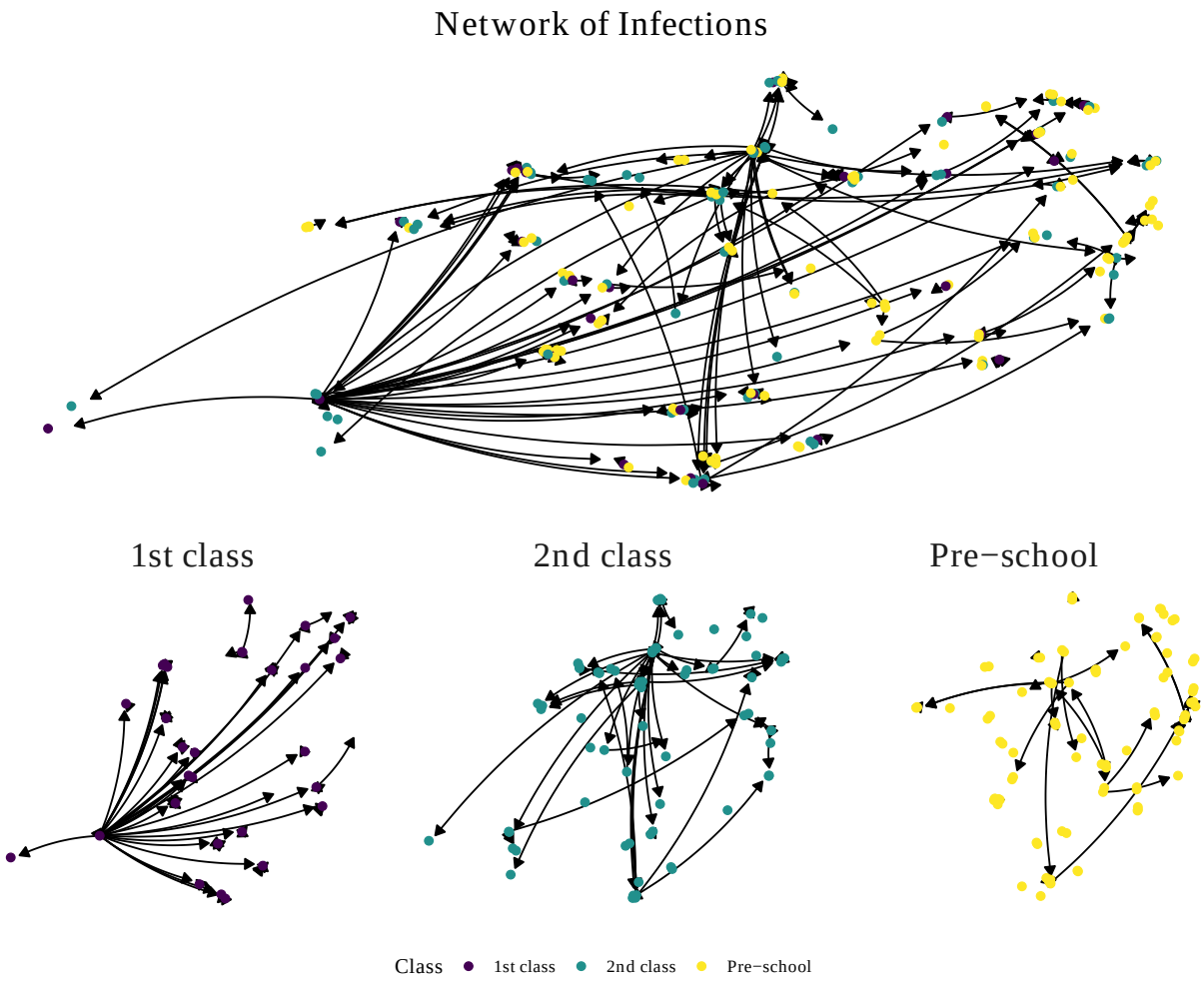

Figure A.3: (Top) Network of infections where nodes (children) are plotted by their household location. The nodes are colored by class, 1st class, 2nd class, or pre-school. (Bottom) Network of infections faceted individually by class. Nodes are still household locations, rescaled.

The number of infections from an infector who is both a sibling and a classmate is 19. Of the 184 children where the purported infector was recorded, 167 ($\sim 90\%$) were infected by either a sibling or direct classmate. Of those who were not infected by a sibling or classmate, all but two were at least as young or younger than their infector, with the median difference of the age of the infector and the infectee being 6 years, meaning the older children in this group were giving the disease to the younger children.

We looked into the group of 21 children who were infected by neither sibling nor classmate. We ruled out close in age but different classes, family name relationships, and siblings of classmates being the cause

of these infections. Overall, 90% of the transmission can be be accounted for by the network structure of the siblings and classmates.

In Figure A.4, we plot the date of the appearance of the measles rash vs. the date of the appearance of the first symptoms, colored by class. A mixture of children from different classes are infected during the first 15 or so days of the epidemic. The disease then seems to infect the 1st class and soon after the 2nd class. The pre-school class seems to become infected last. Also from this graph, we see a strong positive relationship between the initial symptom appearance date and rash appearance date, with the rash appearing on average 3.94 (sample error: 1.74) days after the initial symptoms.

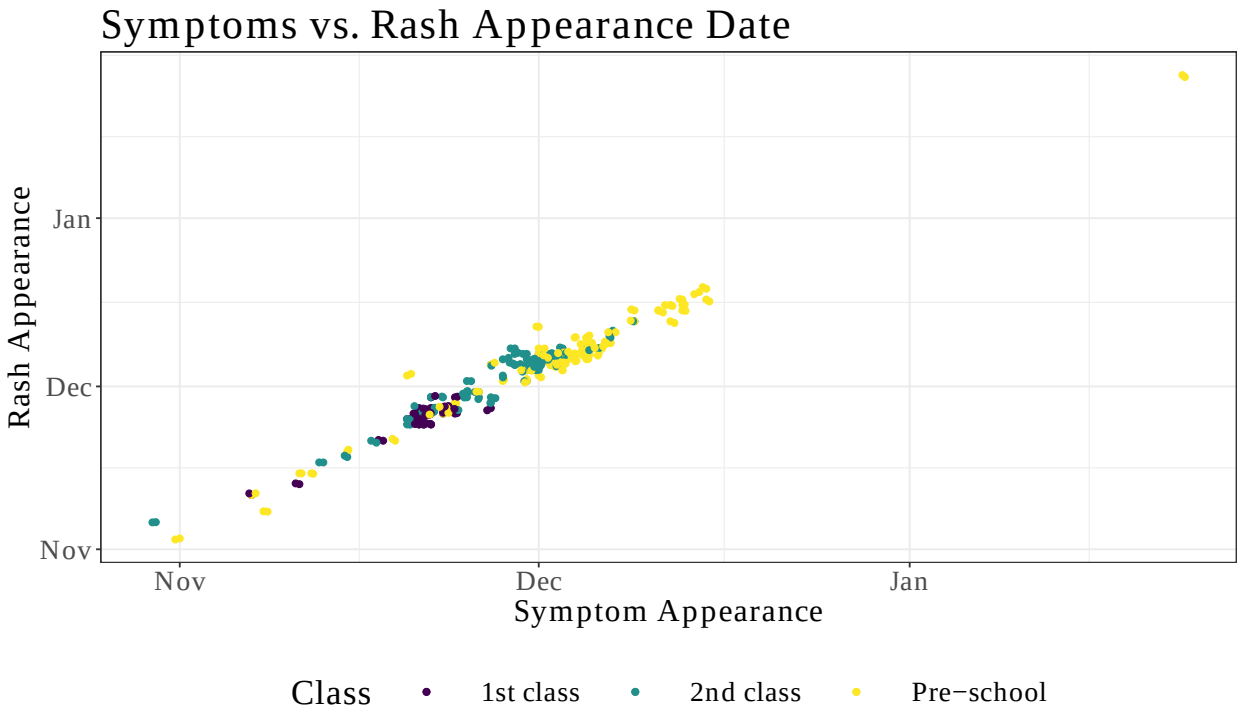

Figure A.4: Plot of rash appearance vs symptom appearance colored by the class.

Vita

Education

2014–2019. PhD in Statistics & Data Science, Carnegie Mellon University

2014-2015. MS in Statistics, Carnegie Mellon University

2010–2014. BS in Mathematical Sciences, Carnegie Mellon University, University and College Honors

Dissertation

July 2019. "Catalyst: Agents of change. Integration of compartment and agent-based models for use in infectious disease epidemiology." Advisor: William F. Eddy. Committee: Joel Greenhouse, Howard Seltman, and Samuel L. Ventura.

Publications

Gallagher, S., Chang, A., and Eddy, W.F. "An analysis of nine ways to estimate $\mathcal{R}_0$ in the SIR model." In preparation, 2019.

Gallagher, S., Frisoli, K., and Luby, A. "Opening up the (surface) of the court in tennis grand slams." In revision, 2019.

Gallagher, S., Richardson, L., Ventura, S.L., and Eddy, W.F. "SPEW: Synthetic Populations and Ecosystems of the World." *Journal of Computational and Graphical Statistics*, 2018.

Selected Presentations and Posters

2018. "Opening up the (court) surface in tennis grand slams." **Honorable Mention** Presentation. Joint work with Kayla Frisoli and Amanda Luby. Carnegie Mellon Sports Analytics Conference. Pittsburgh, PA.

2018. "Catalyst: agents of change." Poster. Advisor: William F. Eddy. Committee: Joel Greenhouse, Howard Seltman, and Samuel L. Ventura. Joint Statistical Meetings. Vancouver, BC.

2017. "Comparing Compartment and Agent-based Models." Presentation. Advisor: William F. Eddy. Committee: Joel Greenhouse, Howard Seltman, and Samuel L. Ventura. Joint Statistical Meetings. Baltimore, MD.

2017. "Generating Synthetic Ecosystems: A Tutorial". **Invited presentation.** Joint work with Lee Richardson, Samuel Ventura, and William Eddy. International Conference on Synthetic Populations. Lucca, Italy.

2016. "Women in Statistics at Carnegie Mellon University." Joint work with Purvasha Chakravarti. Presentation. Women in Statistics and Data Science. Charlotte, NC.

2016. "Statistical Modelling of Infectious Diseases: Influenza and the 'Next Disease.'" Poster. Joint work with Roni Rosenfeld, Ryan Tibshirani, Lee Richardson, Samuel Ventura, and William Eddy. Women in Statistics and Data Science. Charlotte, NC.

2016. "Services for the MIDAS Network: Visualization and Synthetic Ecosystems." Poster. Joint work with Lee Richardson, Samuel Ventura, and William Eddy. MIDAS National Conference. Washington D.C.

2016. "From Forecasting the Flu to Predicting the 'Next' Disease." Poster. Joint work with Roni Rosenfeld, Ryan Tibshirani, Lee Richardson, Samuel Ventura, and William Eddy. UP-STAT. Buffalo, NY.

Honors and Awards

2018. Carnegie Mellon University Sports Analytics Conference Reproducible Paper Competition **Honorable Mention** and **$1,000 award**.

2014, 2018. Gertrude M. Cox Scholarship **Honorable Mention**; ASA Committee on Women in Statistics and the Caucus for Women in Statistics.

2018. Scholarship recipient for the Summer Institute in Statistics and Modeling (tuition and travel stipend). University of Washington, Seattle, WA.

2017. AT&T Labs Graduate Student Symposium Selected Presenter. One of fourteen PhD students out of 79 applicants selected to give a presentation on ongoing research to AT&T researchers in NYC. **Awarded $800** in travel funding.

2016. MIDAS MISSION Public Health **Hackathon Champion** with **$3,000** prize.

2016. UP-STAT **2nd Place Student** Presentation.

2014. Judith A. Resnik Award for Outstanding Women in the Sciences; Carnegie Mellon University.

2013. Phi Beta Kappa Honor Society.

Software

2018-Present. `catalyst`. `R` packaeg for CM-AM pair fitting and diagnostics.

2016-2019. `spew`: `R` package for synthetic ecosystem generation. Lee Richardson, Shannon Gallagher, Samuel L. Ventura, and William F. Eddy.

2017. `spew_dl`: `R` Shiny application to easily browse our synthetic ecosystems produced by `spew`. Shannon Gallagher, Lee Richardson, Samuel L. Ventura, and William F. Eddy.

2016. `spewview`: `R` Shiny application for infectious disease visualization. Shannon Gallagher and Lee Richardson.

Research, Teaching, and Work Experience

2014-2019. **Research Assistant,** Carnegie Mellon University. Generated high-resolution synthetic ecosystem of U.S. and 70+ countries for use in agent-based models for transmission of disease.

2012-2019. **Teaching Assistant**, Carnegie Mellon University. Oversaw lab for 100 students, organized, and led review sessions for a variety of statistics and mathematics classes including Epidemiology, Statistical Computing, Intro to Probability, Advanced Undergraduate Research, Concepts of Mathematics, and Calc 3D.

2015. **Graduate Intern**, PNC. Scraped and analyzed social media data for sentiment analysis. Parallelized code with Hadoop.

Programming Languages

`R` (expert), `julia` (proficient), `html` (intermediate) `Python` (intermediate), `C++` (intermediate), `C` (intermediate), `SQL` (some experience), `jekyll` (some experience)

Professional Service

2018-2019. **PI**. ProSEED/Crosswalk recipient for $1600 to seed a mentorship program across all levels of students within the Stat&DS community.

2017-2019. **Co-President**. Carnegie Mellon University Women in Statistics.

- Organized Women in Data Science Pittsburgh @CMU as an Executive Committee Member. Invited speakers and sponsors, helped organize venue logistics, sent out invitations for attendance, and created the website (2018).
- Maintained Women in Statistics website (2017-2018).
- Organized a seminar by a former CMU PhD student about her experiences as a post-doc at Harvard Biostatistics (2017).
- Organized a panel about applying to graduate school for 30+ undergraduate and masters students (2017).
- Organized dinner with new dean of Mellon College of Science (2017).

2016-Present. **Reviewer**. Statistics in Medicine and Journal of Quantitative Analysis in Sports.

2016-2018. **Co-Organizer**. Pittsburgh use`R`. Organized meet-ups for 30+ members on a variety of topics including cross-language coding and integrating `R` with `github`.

2016-2017. **Judge and volunteer**. Tartan Data Science Cup – 3 to date.

2016-2017. **Vice President**. Carnegie Mellon University Women in Statistics.

2016. **Presenter**. Coding for Girls.

Relevant Course Work

- Machine Learning I & II (**Grad**)
- Statistical Computing (**Grad**)

- Modern Regression (**Grad**)
- Hierarchical Models (**Grad**)
- Multivariate Methods & Data Mining
- Data Matching and Record Linkage
- Advanced Methods for Data Analysis
- Epidemiology

Volunteering

2016-Present. **Family House**. Make meals for families with members in the hospital approximately every other month.

2017-2018. **Stat Help Network**. Hold anonymous "office hours" for graduate students within the Statistics and Data Science Department in order to support students.