Comparing compartment and agent-based models

Shannon Gallagher
Advisor: William F. Eddy

Department of Statistics, Carnegie Mellon University
October 25, 2017

Abstract

Infectious diseases threaten the well-being of society through direct infection of individuals and billions of dollars in collateral damage. As a consequence, statistical modelling of infectious disease plays a critical role in answering important questions about prediction and inference, and additionally, contingency planning. Compartment models (CMs) and agent-based models (AMs) are two common frameworks to disease modelling. Despite the differences between equation-based CMs and simulation-based AMs, researchers have noted substantial similarities between the two frameworks. We intend to combine the two into a "hybrid" framework. We focus on reconciling the statistical differences between CMs and AMs. Specifically, we study a well-known disease framework, Susceptible-Infectious-Recovered, with both a CM-based and AM-based framework. We develop and prove conditions under which these two frameworks have identical statistical behavior. We then extend this equivalence to a large class of CM-AM pairs, which allows for a basis of comparison. Additionally, we examine the relationship between the number of agents and the number of runs in the SIR-AM, which allows for improved computational performance via parallelization. For future work, we propose to extend our current work to all valid compartment models. This will include the development of statistical tests to compare two models to one another in order to measure the differences between them. We will also introduce practical, statistical methods to speed up AM computation time. Finally, we will examine the number of agents required to obtain adequate results to create a statistically justified hybrid model that will simulate a global epidemic.

1 Introduction

Infectious diseases are relevant to modern society. Diseases such as smallpox, polio, and influenza have infected millions of people and induced billions of dollars spent fighting the diseases (Hethcote, 2000). Different generations have recollections of anxieties about polio, HIV, swine flu, and more. Other diseases such as Dengue, Ebola, and Zika have only recently posed threats to the well-being of society (Chan, 2014; Roth et al., 2014; Oliveira Melo et al., 2016). Infectious diseases need to be studied and combated, but the difficulty is in choosing how to do just that.

Two common frameworks used to examine the spread and containment of infectious disease are compartment models (CMs) and agent-based models (AMs). CMs have a long, rich history in epidemiology, whereas AMs represent more recent computer-based methods. CMs and AMs are used to ask and answer the same questions including the typical statistical goals of prediction (e.g. how bad is the flu going to be this year and when?) and inference (e.g. do travel patterns have any effect on the spread of Ebola?). However, as even the *possibility* of an outbreak is quite serious, CMs and AMs are often used to ask a question less seen in statistics, that is, “what-if” scenarios. What if the state issues a travel ban? What if we inoculate the most social individuals within a population? What if everything goes wrong? Investigating prevention strategies such as quarantine and vaccination are important issues in both frameworks. As such, CMs and AMs are tasked to answer important questions which makes deciding between the two all the more important.

Upon first inspection, CMs and AMs seem quite different. CMs are equation-based models that rely on strict assumptions about homogeneity, whereas AMs are simulation-based with heterogeneous agents. Conversely, CMs are expected to run fast whereas AMs may take weeks to run on a computer.

However, researchers have begun to realize that these frameworks are much closer than initially thought. CMs may become quite complex if we partition the population into homogeneous sub-populations, and many AMs combine agents into homogeneous groups to reduce computational time. As some CMs require simultaneous solving of large systems of differential equations, some AMs can run faster than CMs.

As the similarities between CMs and AMs became apparent, researchers began to build hybrid models, which attempted to leverage the advantages of the two while minimizing their disadvantages. However, most of these approaches are *ad hoc* and rarely, if ever, take statistical details such as variance and probability distributions into account.

We examine a well-known epidemic framework, the SIR model, where Susceptible individuals may become infectious and Infectious individuals may become Recovered, in both CM and AM-based frameworks. Specifically, we describe two stochastic models, one from each framework, where the number of individuals in each compartment for each time step, are not only “similar” but equivalent in distribution. We also extend this equivalence to more general CM-AM pairs. This equivalence serves as a basis from which we can then measure how CMs and AMs diverge from one another. We show that by fitting the original SIR model to both frameworks, we obtain the same distribution of epidemic parameters, a concept that may be generalized to test whether an AM and CM or an AM and another AM are similar to one another.

Additionally, we examine the model variance of the two SIR models in question and show that they scale in such a way that the number of agents may be replaced by running the model more times. By decreasing the number of agents and increasing the number of runs, we can improve the total computational time as runs may be parallelized without changing the variance.

We further propose to extend our current work to all deterministic CMs. We will develop statistical tests to determine whether CM-AM pairs or AM-AM pairs are the same in terms of compartment sizes. We will also closely examine the number of agents and number of run trade-off under different assumptions of homogeneity. Ultimately, we will use these concepts to build a statistically justified CM-AM hybrid that switches seamlessly between the two types of frameworks. We will create software for this hybrid model, testing on a large-scale epidemic.

The rest of this document is organized as follows. In Section 2, we describe CMs, AMs, and their uses; explain what work has been done in comparing the two sets of models; and discuss hybrid models that have been attempted. In Section 3, we describe our current work done with the SIR model along with extensions to more general AM-CM pairs. Finally, in Section 4, we develop specific next steps to continue our research. Additionally, in the Appendix A we give simulation results of the results discussed.

2 Previous work

We briefly detail the history of CMs, AMs, comparisons of the two frameworks, and finally combinations of the two in what are known as hybrid models.

2.1 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 moves through a population. Perhaps the most well known CM is the SIR model, introduced by Kermack and McKendrick (1927), which stands for susceptible, infectious, and recovered, respectively. 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: homogeneity and 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, the assumption of homogeneity is highly controversial.

Compartment models within the SI framework can be as simple the SI model or can be made 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 actual 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. Bayesian approaches have also been attempted such as the one described in 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 meaning smaller variance. 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, 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.

2.2 Agent-based models

Falling under the broader class of "simulations," agent-based models (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 split 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 vs. stochastic interactions. In Conway's Game of Life, the states of agents in an AM are completely determined given their initial state. 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.

In the field of infectious disease epidemiology, AMs, sometimes called Individual Level Models (ILMs), as agent often has another meaning in this field, are typically used to model the spread of infectious disease (Longini et al., 2004; Grefenstette et al., 2013). AMs in this field have been used for prediction, inference, and study of hypothetical prevention strategies (Eubank et al., 2010; Bajardi et al., 2011; Liu et al., 2015; Wang et al., 2016). Typically in these models, agents are assumed to be non-random as are the environments, with the only variation arising through transference of a disease through activities of agents. Generally, variance is reported through simulation results accumulated by running the model hundreds of times, if at all.

AMs are a step away from the homogeneity of CMs as agents may be quite diverse. Although a CM may include heterogeneous information by adding numerous compartments, such as through a model of AIDS posited by Anderson et al. (1986), it is typically very tedious to do so. As such, AMs allow for an easier way to incorporate heterogeneity into a model.

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. (2015) 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.

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), and we adapt their notation of AMs here. That work, however, focuses on modelling the underlying probability of evolving from one state to another rather than the statistical properties of the AM.

To summarize, the shortcomings AMs are two-fold 1) aligning the model to reality and 2) computational time. Wallentin and Neuwirth (2017) describe this as the computational-predictive trade-off.

2.3 Comparing 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. Moreover, although 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).

2.4 Hybrid models

Some modellers 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. Continuing the global versus local effects, Nguyen et al. (2008) examine two "patches" of an environment. 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 under a certain number of infected individuals and then switches to 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.

The idea of hybrid models is generally well-received but the details, statistical and otherwise, on how to create such a model are currently lacking.

3 Current work: the SIR framework

In this section, we examine the relationship between CMs and AMs within the context of the SIR framework. After describing the SIR model, we present two stochastic versions: a CM and AM. We show that these models are very similar, and in some ways, exactly the same. We then extend this equivalence to general CM-AM pairs.

In the SIR model, individuals are partitioned into one of three groups, susceptible, infectious, or recovered. Susceptible individuals may become infectious and infected individuals may become recovered over time. Both CMs and AMs may arise from the SIR framework, depending on what assumptions are made about the transitions.

3.1 SIR: the CM approach

The deterministic CM of the SIR framework dates back to the work of Kermack and McKendrick (1927). Their model follows the CM framework by assuming homogeneity of individuals within compartments and by describing the transitions between compartments through a set of equations. For a fixed population N and a given unit of time, the time to infection is expected to be β^{-1}, and the recovery time of infected individuals is expected to be γ^{-1}. Thus, if the initial number of individuals in each compartment is known $(S(0), I(0), R(0))$, then the entire model is specified through β and $\gamma \in (0, 1]$. This is expressed by the following set of difference equations shown in Equation (1),

$$\begin{cases} \frac{\Delta S}{\Delta t} = -\frac{\beta IS}{N} \\ \frac{\Delta I}{\Delta t} = \frac{\beta IS}{N} - \gamma I \ . \\ \frac{\Delta R}{\Delta t} = \gamma I \end{cases} \tag{1}$$

Notably, in this model both $S(t)$ and $R(t)$ are monotonic functions where $S(t)$ is non-increasing and $R(t)$ is non-decreasing. No such restriction is placed on $I(t)$.

Typically, the SIR-CM is treated as deterministic, but random movement among compartments may be incorporated. Denote $\hat{S}(t), \hat{I}(t)$, and $\hat{R}(t)$ to be the observed number in each compartment and $S(t), I(t)$, and $R(t)$ to be the true, underlying model at time t. A reasonable model is given by the following:

$$\begin{aligned} \hat{S}(t+1) &= \hat{S}(t) - s_t \\ \hat{I}(t+1) &= N - \hat{S}(t+1) - \hat{R}(t+1), \\ \hat{R}(t+1) &= \hat{R}(t) + r_t \end{aligned} \tag{2}$$

with $\hat{S}(0)$, $\hat{I}(0)$, and $\hat{R}(0)$ known. The random terms in Equation (2) are random variables. Here we choose them to have the following distributions,

$$\begin{aligned} s_t &\sim \text{Binomial}\left(\hat{S}(t), \frac{\beta I(t)}{N}\right) \\ r_t &\sim \text{Binomial}\left(\hat{I}(t), \gamma\right) . \end{aligned} \tag{3}$$

One reason that this model was chosen is that, on average, it follows the shape of the original SIR model. Another reason why this model was chosen is due to monotonic conditions placed on the S and R components, non-increasing and non-decreasing, respectively.

3.2 SIR: the AM approach

We can also create an AM-model for the the SIR frameework. AMs include dynamic agents $a_n(t)$ for $n = 1, 2, \ldots N$ along with a forward operator which updates the state of the agents from one step to the next. The variable of interest is the aggregate total of the agents that belong to the compartment k, $\hat{X}_k(t)$ for $k = 1, 2, \ldots, K$. The following AM relies on an underlying, deterministic SIR-CM. Thus, we assume a fixed population N that follows the underlying model $S(t)$, $I(t)$, and $R(t)$ with known initial states $S(0)$, $I(0)$, and $R(0)$, respectively. The AM is initialized to match that of the underlying SIR model, e.g $\hat{X}_1(0) = S(0)$, $\hat{X}_2(0) = I(0)$, and $\hat{X}_3(0) = R(0)$. The forward operator is dependent on the current compartment of the agent in question. An agent in a given state has a probability to move to the next state based on the underlying SIR-CM,

$$a_n(t+1) = \begin{cases} a_n(t) + \text{Bernoulli}\left(\frac{\beta I(t)}{N}\right) & \text{if } a_n(t) = 1 \\ a_n(t) + \text{Bernoulli}\left(\gamma\right) & \text{if } a_n(t) = 2 \\ a_n(t) & \text{otherwise} \end{cases} \cdot \tag{4}$$

Here, compartments 1, 2, and 3 refer to the S, I, and R compartments, respectively. Thus, Equation (4) fully describes an AM within the SIR framework.

The variables of interest are the total number of agents in each compartment, $k \in \{S, I, R\}$ at each time step $t \in \{1, 2, \ldots, T\}$ and where $\mathcal{I}$ is the indicator function,

$$\hat{X}_k(t) = \sum_{n=1}^{N} \mathcal{I}\left\{a_n(t) = k\right\}. \tag{5}$$

Note the expected value of $\hat{X}_k$ for each of the compartments is that of the respective compartment in the underlying deterministic SIR model.

3.3 SIR: CM vs. AM

Many researchers have mentioned that sometimes AMs and CMs act the same and sometimes differently, but we give conditions where not only the AM and CM's expected values match those of the underlying SIR model but also are the same in distribution for the number of individuals in each compartment. The models chosen for the CM described by Equations (2) and (3) and the AM in Equation (4) were chosen precisely because under these conditions, these models are not only similar, but exactly the same.

Theorem 1. *Fix an underlying SIR model, $S(t), I(t)$, and $R(t)$ with known $S(0)$, $I(0)$, and $R(0)$. Let the CM be as in Equations (2) and (3) and the AM in Equation(4). Then for all $t \in \{1, 2, \ldots, T\}$,*

$$\left(\hat{S}(t), \hat{I}(t), \hat{R}(t)\right) \overset{d}{=} \left(\hat{X}_S(t), \hat{X}_I(t), \hat{X}_R(t)\right) \tag{6}$$

Proof. The initial conditions are designed to be the same in each model. Noting that the Binomial in the CM model can be thought of a sum of independent Bernoullis, the claim subsequently follows. □

3.4 Equivalent General CMs and AMs

The idea of equivalent CMs and AMs generalizes to other models besides the deterministic SIR model. We can create equivalent CM-AM pairs for any given deterministic CM given by a series of difference equations, given the equations satisfy the law of mass action (in order to specify valid probabilities). We first give the conditions for the stochastic CM.

Let a deterministic CM with K discrete compartments be given by a series of K difference equations,

$$\frac{\Delta X_k^{CM}}{\Delta t} = \sum_{i=1}^{K} D_{ik}(t) - \sum_{j=1}^{K} D_{kj}(t)$$

along with a set of initial values. Here $D_{ij}(t) \geq 0$ represents the movement of individuals from compartment i to compartment j from time t to $t+1$. The stochastic CM is given by the following which includes Z_{ij} which represents the jth entry of vector Z_i,

$$Z_i \sim \text{Multinomial}\left(\hat{X}_i, (p_{i1}(t), \ldots, p_{iK}(t))\right) \text{ for } i = 1, 2, \ldots, K \tag{7}$$

$$\hat{X}_k^{CM}(t+1) = \hat{X}_k^{CM}(t) + \sum_{i=1}^{K} Z_{i,k} - \sum_{i=1}^{K} Z_{k,i} \text{ for } k = 1, 2, \ldots, K$$

where $p_{ij}(t) = \frac{D_{ij}(t)}{X_i(t)}$ when $i \neq j$ and $p_{ii}(t) = 1 - \sum_{k \neq i} p_{ik}$. In words, the random variables Z_i are multinomial draws where entry Z_{ij} is the number of individuals moving from compartment i to compartment j. Then the number of new individuals in compartment k is equal to the old number plus the new individuals moving in and minus individuals moving out.

Similarly, we can create a corresponding AM for this deterministic CM. Given the correct initial values in each compartment, for an agent $a_n(t)$, $n = 1, 2, \dots, N$, the forward operator for $t > 0$ is

$$a_n(t+1) = j \text{ with probability } p_{kj}(t) \text{ if } a_n(t) = k. \tag{8}$$

That is, an agent currently in state k has probability $p_{kj}(t)$ of moving into state j. The aggregate total in each compartment k is given by

$$\hat{X}_k^{AM}(t) = \sum_{n=1}^{N} \mathcal{I}\left\{a_n(t) = k\right\}.$$

Then, these CM-AM pairs are equivalent.

Theorem 2. *Fix an underlying deterministic CM with K compartments $X_1, X_2, \dots, X_K$ with known initial values and differential (difference) equations. Let the stochastic CM be as in Equation (7) and the AM in Equation (8). Then for all $t \in \{1, 2, \dots, T\}$,*

$$\left(\hat{X}_1(t), \hat{X}_2(t), \dots, \hat{X}_K\right)^{CM} \stackrel{d}{=} \left(\hat{X}_1(t), \hat{X}_2(t), \dots, \hat{X}_K(t)\right)^{AM} \tag{9}$$

Proof. The initial conditions are designed to be the same in each model. Noting that the Multinomial draws in the CM model can be thought of a sum of independent Multinomial draws of size 1, the claim subsequently follows. □

3.5 Comparing models

AMs rarely fully consider statistics in their methods. For example, simulations are run over entire populations rather than sub-samples. Thus one practical result of considering statistics in AMs is the ability to substantially reduce computational time through either running a CM instead of an AM or combining multiple AMs together. We first need to be able to compare AMs to CMs and AMs to AMs in order to decide when they are "similar enough." One way to do this is to fit a deterministic SIR model and compare the resulting estimated $\hat{\beta}$ and $\hat{\gamma}$ parameters from the different models.

Example 3.1 (SIR: AM & CM: simulations)**.** We first simulate the stochastic SIR model for both the SIR-CM and SIR-AM with $N = 1000$, $\beta = .10$ and $\gamma = .03$. We then fit a deterministic SIR model to each of simulations. We do so by using the initial conditions as in the observed model and jointly minimizing the number of susceptible and infected (and thus the recovered),

$$\mathcal{L}(\beta, \gamma) = \sum_{t=0}^{T} \left(S_{obs}(t) - S(t; \beta, \gamma)\right)^2 + \left(I_{obs}(t) - I(t; \beta, \gamma)\right)^2. \tag{10}$$

Repeating for 5000 draws from the CM and AM described previously, we obtain 5000 pairs of β and γ estimates, the results of which are plotted as a contour in Figure 1. Not surprisingly, the distributions of the pairs from the CM and AM seem to be indistinguishable from one another.

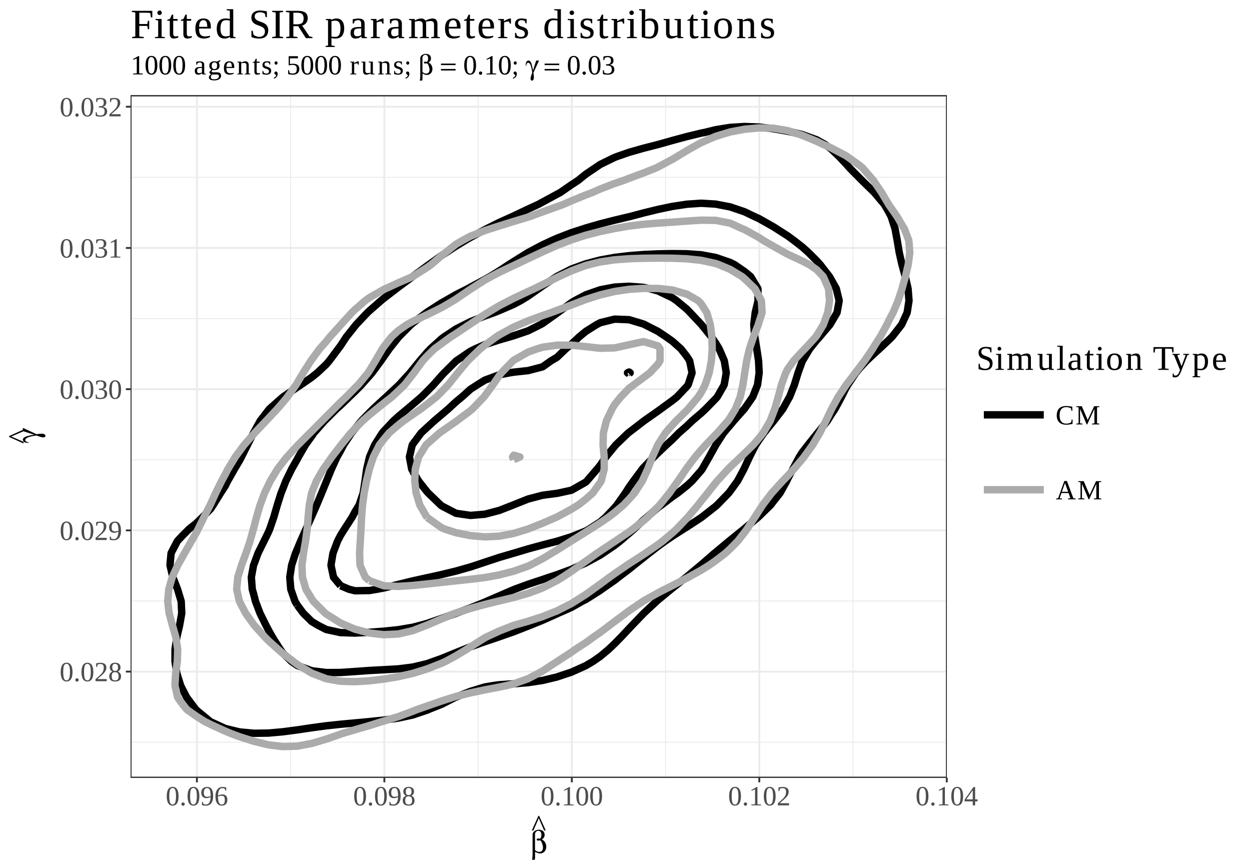

Figure 1: Contour plot of $(\hat{\beta}, \hat{\gamma})$ pairs from 5000 draws from SIR models fit to both the CM (black) and AM (gray) draws described in Equations (2) and (3) and Equation (4), respectively.

3.6 Number of agents vs. number of runs

A key advantage to the AM framework is the ability to run a model with fixed initial parameters a number of times. These runs are independent of one another, which means the process can be completed in parallel. Thus, if we can determine the relationship between the number of runs and the number of agents and their resulting effects on the model, then we can improve computational time via parallelization without compromising information about the variance or the distributions of the different compartments.

Specifically, we want to look at the distribution of $\frac{1}{L}\sum_{\text{runs } \ell} \frac{\hat{X}_k(t)}{N}$, where it is implicitly assumed that $\hat{X}_k(t)$ may change with each run with L total runs. Clearly,

$$\begin{aligned} E\left[\frac{1}{L}\sum_{\text{runs } \ell} \frac{\hat{X}_k(t)}{N}\right] &= E\left[\frac{\hat{X}_k(t)}{N}\right] \\ V\left[\frac{1}{L}\sum_{\text{runs } \ell} \frac{\hat{X}_k(t)}{N}\right] &= \frac{1}{LN^2} V\left[\hat{X}_k(t)\right], \end{aligned}$$

and so the relationship between the number of agents and runs is dependent on how $\hat{X}_k(t)$, the total number of agents in state k at time t, varies as a function of N.

3.6.1 SIR-AM: the number of agents and runs in an AM

Let there be two SIR-AMs with $\hat{X}_{k_1}(t)$ and $\hat{X}_{k_2}(t)$ as the same compartment type (e.g. S, I, or R) of an underlying, deterministic SIR-CM with parameters β and γ with total population, N_1 and N_2, respectively and with initial values such that $\frac{\hat{X}_{k_1}(0)}{N_1} = \frac{\hat{X}_{k_2}(0)}{N_2}$. That is, the only thing that is different between the two models is the population size. These AMs follow the the form as described in Equation (4). First, note that $\frac{\hat{X}_{k_1}(t)}{N_1} = \frac{\hat{X}_{k_2}(t)}{N_2}$ for all t. Then it follows that

$$E\left[\frac{1}{L_1}\sum_{\text{runs } \ell} \frac{\hat{X}_{k_1}(t)}{N_1}\right] = E\left[\frac{1}{L_2}\sum_{\text{runs } \ell} \frac{\hat{X}_{k_2}(t)}{N_2}\right].$$

The variance and covariance are calculated for the SIR compartments using the law of total covariance of the model described in Equation (2). They are, setting $p_t = \frac{\beta I(t)}{N}$:

$$\begin{aligned} V\left[\hat{S}(t+1)\right] =& S(t)(1-p_t)p_t + (1-p_t)^2 V\left[\hat{S}(t)\right] \\ V\left[\hat{I}(t+1)\right] =& V[\hat{S}(t+1)] + V[\hat{R}(t+1)] - 2\text{Cov}[\hat{S}(t+1), \hat{R}(t+1)] \\ V\left[\hat{R}(t+1)\right] =& I(t)\gamma(1-\gamma) + (1-\gamma)^2 V\left[\hat{R}(t)\right] + \gamma^2 V[\hat{S}(t)] \\ & - 2\gamma(1-\gamma)\text{Cov}\left(\hat{S}(t), \hat{R}(t)\right) \\ \text{Cov}\left[\hat{S}(t+1), \hat{R}(t+1)\right] =& -\gamma(1-p_t)V[\hat{S}(t)] + (1-\gamma)(1-p_t)\text{Cov}[\hat{S}(t), \hat{R}(t)]. \end{aligned} \tag{11}$$

As a result of Equation (11),

$$V[\hat{X}_{k_2}(t)] = \frac{N_2}{N_1} V[\hat{X}_{k_1}] \text{ for } k \in \{S, I, R\}$$

if $V[\hat{X}_{k_1}(0)] = V[X_{k_2}(0)] = 0$. This result lets us simply compare the variance of SIR-AMs with different population sizes.

Assume we have SIR-AM 1 with N_1 agents and L_1 runs and known initial values and SIR-AM 2 with N_2 agents and L_2 runs, both with given parameters β and γ. Additionally, assume $\hat{X}_{k_2}(0) = \frac{N_2}{N_1}\hat{X}_{k_1}(0)$, that is, the initial values in each compartment have the same proportion as the other AM. Then

$$\begin{aligned} \frac{V\left[\frac{1}{L_1}\sum_{\text{runs } \ell} \frac{\hat{X}_k(t)}{N_1}\right]}{V\left[\frac{1}{L_2}\sum_{\text{runs } \ell} \frac{\hat{X}_{k_2}(t)}{N_2}\right]} &= \frac{L_2 N_2^2}{L_1 N_1^2} \cdot \frac{V[\hat{X}_{k_1}(t)]}{V[X_{k_2}(t)]} \\ &= \frac{L_2 N_2}{L_1 N_1}. \end{aligned} \tag{12}$$

If we set the variances equal to one another then $L_1 N_1 = L_2 N_2$. This result lets us freely exchange runs with agents without changing the variance. Thus, by reducing the number of agents and increasing the number of runs we can effectively improve the effective computational time of the model without compromising information about the variance. As an example, Figure 2 shows the variance of the compartment size averaged over both the agents and the time steps for two separate simulations of the SIR-AM. Simulation 1 consisted of $N = 100$ agents and $L = 4$ runs which were run simultaneously over 4 cores. Simulation 2 consisted of $N = 400$ agents and $L = 1$ run. Both sets of simulations were executed 100 times and their sample average variance is plotted. We see the variance is approximately the same although simulation 1 took 3:30 minutes to run and simulation 2 took 4:05 minutes to run.

4 Next steps

Future work focuses on investigating when and how AMs and CMs diverge from one another and the development statistical tests to verify the divergence. Ultimately, we hope to construct a statistically justified hybrid model. Specifically, we will examine

- **Extending current work to run on FRED.** We will examine the concept of creating increasingly diverse agents empirically using the existing AM FRED (Grefenstette et al., 2013). FRED follows a SEIR framework but allows for flexible parameter inputs. First, we will modify the SEIR equations to exclude the "E" compartment and thus use the SIR framework. We will then first replicate the results of this document, that is we will examine the spread of disease in a completely homogeneous population. From there, we will begin to add heterogeneity to our agents. We will first partition our agents by splitting one agent feature into two groups, which will make "parallel" SIR models and the previous results should still apply. We will then partition the agents on two features with two groups for each feature. This will allow for heterogeneous mixing of the agents. We will change the infectivity and recovery rates for the different groups and examine how the disease spreads. In all these experiments, we will focus on the statistical details such as variance and probability distributions in addition to the mean results.
- **Examination of foundational assumptions.** In our proposal work, we created AM-CM pairs using the assumption of independent agents. However, we know

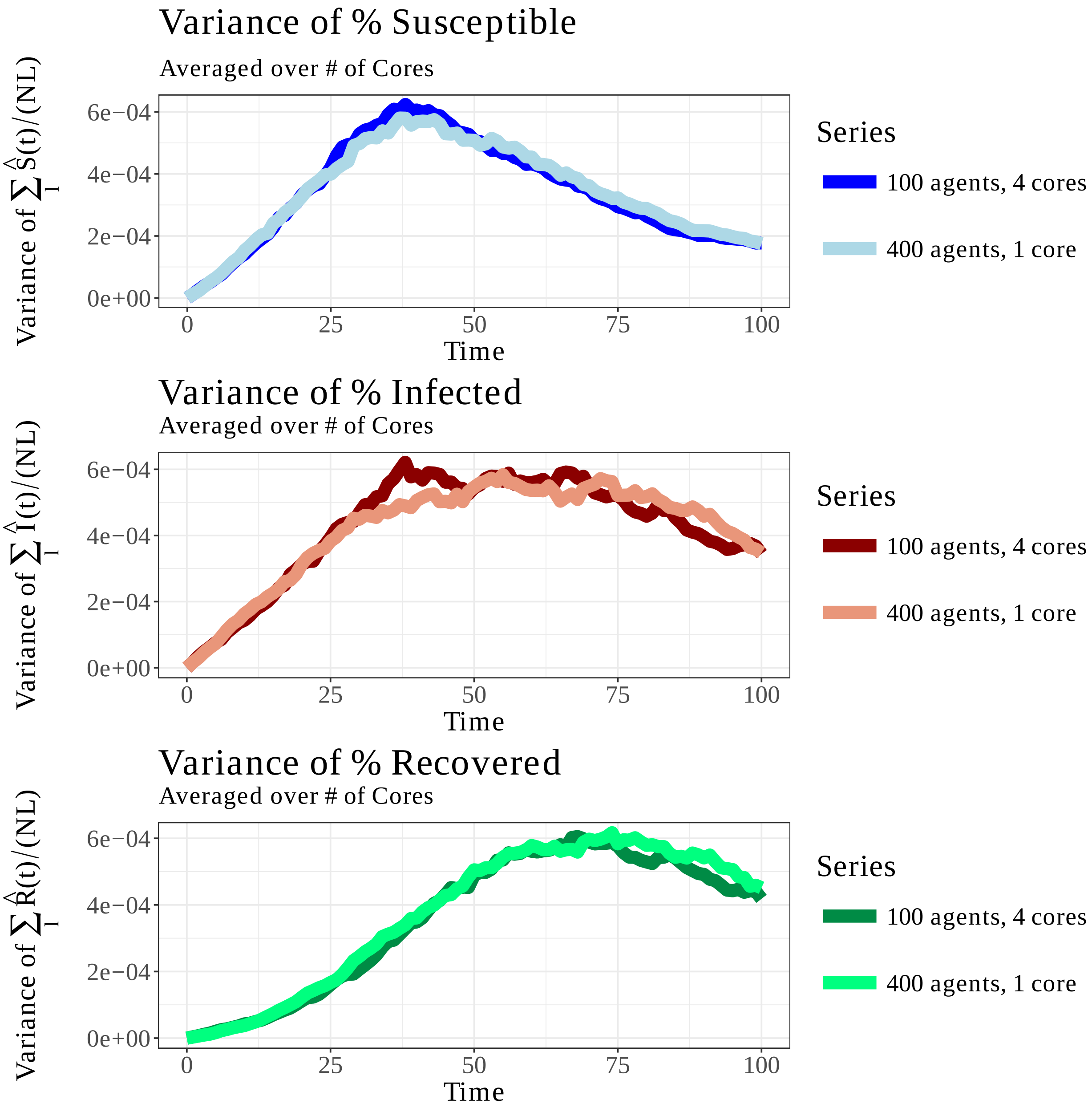

Figure 2: Variance of the compartments for two simulations, both with $\beta = .1$ and $\gamma = .03$. Simulation 1: $N = 100$ agents, $L = 4$ runs. Simulation 2: $N = 400$ agents, $L = 1$ run. Each simulation was run 100 times. Simulation 1 ran for 3:30 minutes and simulation 2 for 4:05 minutes as simulation 1 was parallelized over 4 cores.

this is a sufficient but not necessary condition, as we will show through using a series of distinct groups of agents that are mutually independent. We would like to further examine the fundamental assumptions of CMs and their corresponding AMs including that of independence of agents, homogeneity of agents, and the law of mass action. We would like to examine the conditions necessary for each of these assumptions to create equivalent AM-CM pairs. Additionally, a simple parameter to change in our equivalent AM-CM pairs is the probability of transition from compartment i to j from time t to $t+1$, $p_{ij}(t)$,. It would be interesting to examine the effect p_t has on the variance of the resulting model under different assumptions, such as a Beta prior.

- **Statistical tests to compare AMs to CMs.** We would like to form tests to decide when it is justified to switch from an AM to a CM. Points to consider are when in time a switch is justified, what is the effect on the variance, how do we obtain parameters for the CM, and whether prevention strategies will be effected and how so. This is the first step in creating a statistically justified hybrid model. We will first compare fitted epidemiological parameters such as β and γ using statistical tests to differentiate AMs from one another. From there, we will explore higher dimensional summary statistics. We will explore statistical properties of similarity scores used to quantitatively compare AMs and CMs.

- **Examination of methods on real data.** We will run our methods on real infectious disease data such as the recent Ebola outbreak in Western Africa or possibly another disease. We examine CMs and AMs in this setting and whether the resulting fitted $\hat{\beta}$ and $\hat{\gamma}$ parameters are effected in each model.

- **Software for a hybrid model.** We will demonstrate a hybrid model computationally with focus on disease applications. We will introduce a global epidemic to the populations of Synthetic Populations and Ecosystems of the World (Gallagher et al., 2017), which will apply a more AM-based approach in more detailed and thus heterogeneous synthetic ecosystems such as the United States and more of a CM-based framework in less detailed synthetic ecosystems.

References

Abbey, H. (1952). An examination of the reed-frost theory of epidemics. *Human biology*, 24(3):201.

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

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.

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

Anderson, R. and May, R. (1992). *Infectious Diseases of Humans*. Oxford: Oxford University Press.

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.

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.

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.

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.

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

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.

Chan, M. (2014). Ebola Virus Disease in West Africa No Early End to the Outbreak. *New England Journal of Medicine*, 371(13):1183–1185. PMID: 25140856.

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.

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

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).

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

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.

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.

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

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.

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.

Gallagher, S., Richardson, L., Ventura, S. L., and Eddy, W. F. (2017). SPEW: Synthetic Populations and Ecosystems of the World. *To appear*.

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):10631076.

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.

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.

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

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

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.

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.

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.

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.

Liu, F., Enanoria, W. T., Zipprich, J., Blumberg, S., Harriman, K., Ackley, S. F., Wheaton, W. D., Allpress, J. L., and Porco, T. C. (2015). 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).

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.

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

Nguyen, N. D., Drogoul, A., and Auger, P. (2008). *Methodological Steps and Issues When Deriving Individual Based-Models from Equation-Based Models: A Case Study in Population Dynamics*, pages 295–306. Springer Berlin Heidelberg, Berlin, Heidelberg.

Oliveira Melo, A. S., Malinger, G., Ximenes, R., Szejnfeld, P. O., Alves Sampaio, S., and Bispo de Filippis, A. M. (2016). Zika virus intrauterine infection causes fetal brain abnormality and microcephaly: tip of the iceberg? *Ultrasound in Obstetrics & Gynecology*, 47(1):6–7.

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.

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.

Roth, A., Mercier, A., Lepers, C., Hoy, D., Duituturaga, S., Benyon, E., Guillaumot, L., and Souares, Y. (2014). Concurrent outbreaks of Dengue, Chikungunya and Zika virus infections-an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012-2014. *Euro Surveill*, 19(41):20929.

Rvachev, L. A. and Longini, I. M. (1985). A mathematical model for the global spread of influenza. *Mathematical Biosciences*, 75(1):3 – 22.

Scheffer, M., Baveco, J., DeAngelis, D., Rose, K., and van Nes, E. (1995). Superindividuals a simple solution for modelling large populations on an individual basis. *Ecological Modelling*, 80(2):161 – 170.

Schelling, T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1.

Smith, L., Beckman, R., and Baggerly, K. (1995). *TRANSIMS: Transportation analysis and simulation system.*

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.

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.

Wang, Z., Bauch, C. T., Bhattacharyya, S., dOnofrio, A., Manfredi, P., Perc, M., Perra, N., Salathé, M., and Zhao, D. (2016). Statistical physics of vaccination. *Physics Reports*, 664:1–113.

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.

Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media, Inc.

A Simulation Results

Example A.1 (Equivalent SIR CM & AM: simulations)**.** The CM is drawn from the model in Equations (2) and (3) and the AM as in Equation (4). In the underlying SIR model, $\beta = 0.10$ and $\gamma = 0.03$ and $N = 1000$ with $S(0) = 950$, $I(0) = 50$ and $R(0) = 0$, and draw 1000 instances from both the CM and the AM.

The results are displayed in Figures 3-4. From Figure 3, we see that the draws overlaid on one another for both the AM and the CM seem to have the same shape which is indicative of having the same distribution for the observed S, I, and R curves. In Figure 4, the average value at each time point for the observed S, I , and R curves are plotted for both the AM and CM. There is no distinguishable difference between the two sets of curves, indicating that the mean values for each time point and curve are equal.

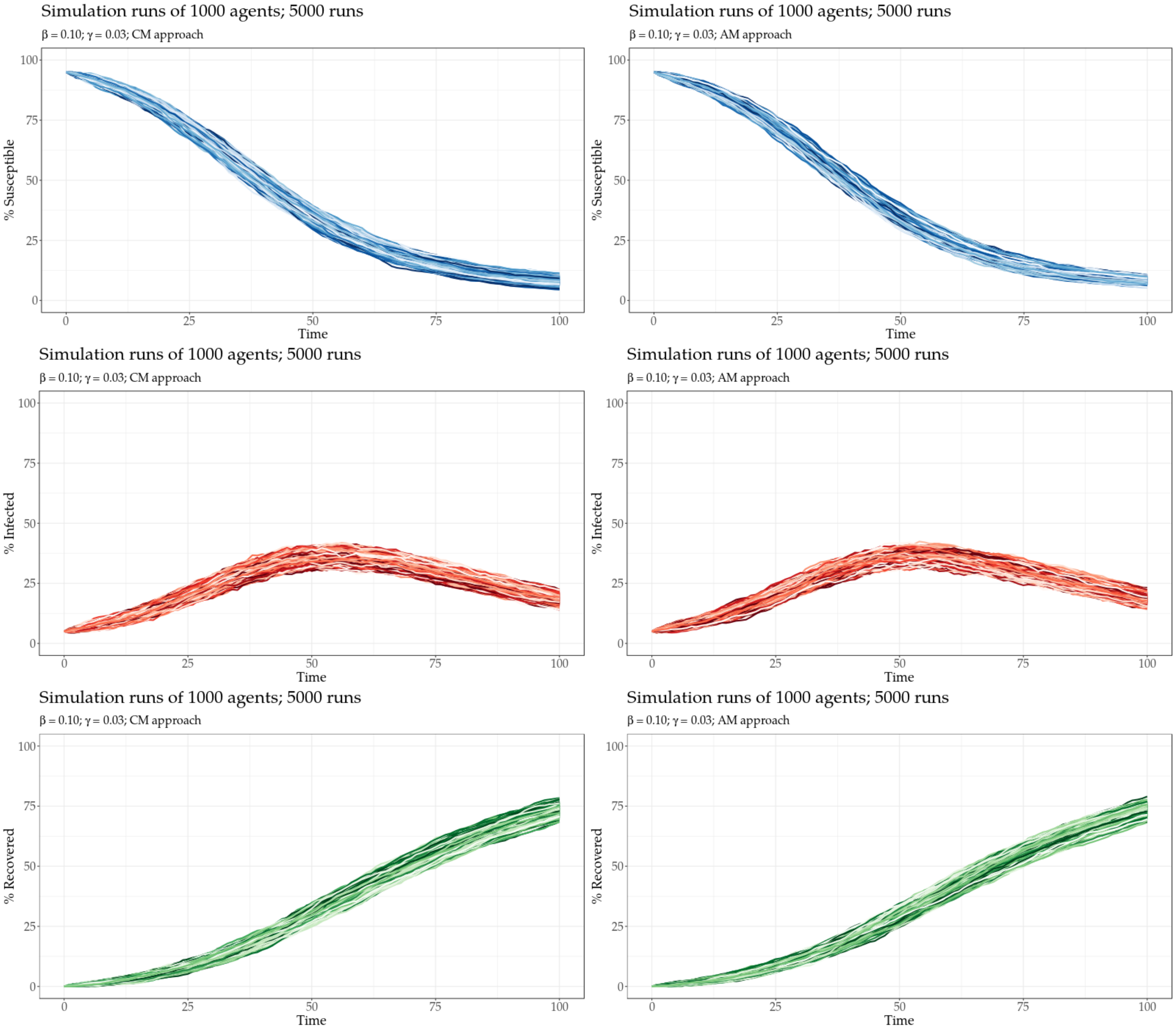

Figure 3: 100 draws from the CM (left) and AM (right) for the observed S, I, and R curves described in Equations (2) and (3) and Equation(4), respectively.

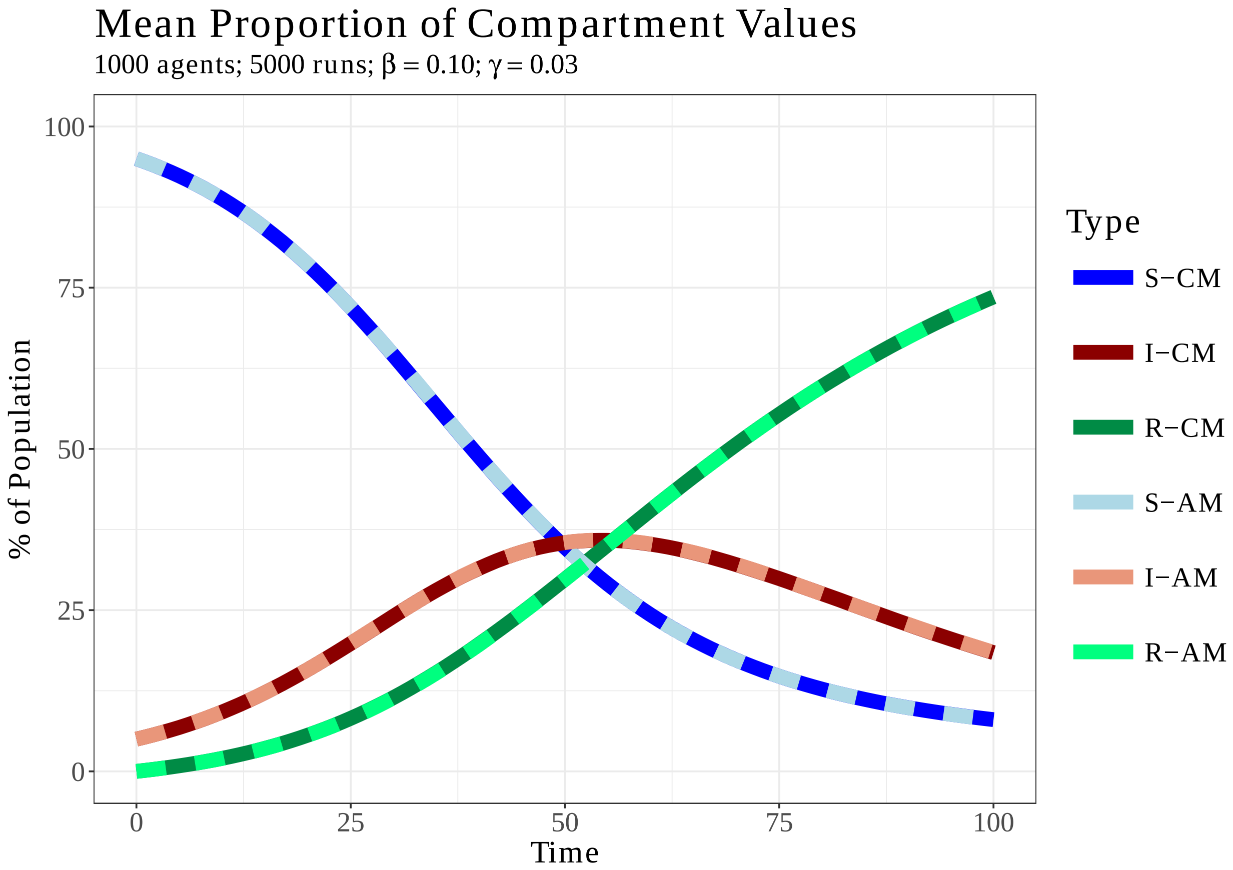

Figure 4: The average value for each time point of 100 draws from the CM and AM for observed the S, I, and R curves described in Equations (2) and (3) and Equation (4), respectively. The two sets of lines completely overlap.