2021

Addressing the Ecological Fallacy with Lagrangian Inference

Michael Schwob
schwom1@unlv.nevada.edu

Addressing the Ecological Fallacy with Lagrangian Inference

By

Michael Richard Schwob

Honors Thesis submitted in partial fulfillment

for the designation of Research and Creative Honors

Department of Mathematical Sciences

Thesis Advisor: Dr. Kaushik Ghosh

Thesis Committee: Dr. Bryan Bornholdt & Dr. Hokwon Cho

College of Sciences

University of Nevada, Las Vegas

May, 2021

Acknowledgments

I would like to thank my thesis adviser (Dr. Kaushik Ghosh) and committee (Dr. Bryan Bornholdt and Dr. Hokwon Cho) for providing feedback throughout. I would also like to thank Dr. Mevin Hooten for bringing me into the world of Lagrangian inference and my parents for bringing me into the world.

Abstract

Most epidemiologists elect to use statistical models that use population-level data to make inference on the spread of some virus or disease. This has become commonplace in the fields of epidemiology and biostatistics since most data used to construct and verify epidemic models are recorded at the population-level. Obtaining inference from a population-level model may be beneficial in studying the spread of disease in a homogeneous population, but the use of such models to describe a heterogeneous population results in inadequate inference. The inaccuracy of these models is further amplified when one tries to make individual-level inference from these population-level models. This thesis argues for the adoption of individual-level (Lagrangian) inference when attempting to obtain inference for an individual or a heterogeneous population. To support this argument, an example of the ecological fallacy is provided and an epidemic agent-based model is delineated to analyze the SARS-CoV-2 outbreak on the Diamond Princess cruise ship. To aid in simulation, a surrogate model is discussed that interpolates analyses for the computationally expensive agent-based model. Finally, the extension of such a method to larger data sets, such as Clark County, Nevada, is considered.

Contents

1 Introduction

Due to the nature of data collection in epidemiology, most disease data sets present data at the population level rather than the individual level. This population-level aggregation of data can be attributed to the ethics of sharing biomedical data between organizations and institutions. As a result of such structured data, epidemiologists mainly build models from a population (or Eulerian) perspective. On large data sets, such models typically run much quicker than models that adopt an individual-level (or Lagrangian) approach, further pushing their adoption within the field of epidemiology.

Epidemiologists aim to extract accurate information from their models. The resulting inference is made for the population of interest in a study. For most epidemiological studies, the population of interest is well-represented in the sample that is observed. Inference is then relayed from the epidemiologists to the public, pharmaceutical companies, or health institutions. Sound epidemiological practice requires that the population of interest is contained within the study's sample, which results in more accurate and personalized inference. When the population of interest and study sample coincide, the resulting inference can be applied to each member of the population. However, if the population of interest is heterogeneous, the resulting population-level inference may not be applicable to every member in the population. Thus, an Eulerian model may not be the most robust model for heterogeneous populations.

An ecological fallacy is the misinterpretation of statistical data that occurs when inferences about the nature of individuals are deduced from inferences about the group to which those individuals belong. This fallacy claims that one cannot obtain accurate inference for an individual from a group. When epidemiologists attempt to make individual-level inference from a heterogeneous population with an Eulerian model, they are committing an ecological fallacy. In the case that individual-level inference is desired, Lagrangian models should be used. This is certainly the case for an ongoing global pandemic.

SARS-CoV-2 data sets aggregate individuals' data to some population level, such as the

'county', 'state', 'country', or 'global' level. With SARS-CoV-2, the public is interested in their infection or death probability. That is to say, each individual is interested in their own probabilities, and maybe the probabilities of some loved ones —not that of the entire population. The public is interested in individual-level inference; however, the information they are receiving is generated from population-level models. This thesis argues for the adoption of Lagrangian models when individual-level inference is desired.

The remainder of this thesis is organized as follows. A thorough literature review is provided in Section 2. In Section 3, a novel simulation demonstrates an ecological fallacy, which encourages the use of Lagrangian methods for epidemiological inference. An example of a suitable Lagrangian model is delineated in Section 4, and its analysis of the Diamond Princess cruise ship data is shared in Section 5; the methodology and application were both published in one of my recent papers with Dr. Mevin Hooten and Dr. Christopher Wikle [28]. Section 6 reveals the possible extension and anticipated problems of using this framework for larger data sets, such as for Clark County, Nevada. Finally, some concluding remarks are given in Section 7.

2 Literature Review

Recently, there has been a small wave of Lagrangian model adoption in epidemiology. With such models, epidemiologists are able to adjust their parameters and make accurate individual-level inference within a heterogeneous population. With their individual-level inference in hand, they can then scale their inference up to the population level.

Agent-based models (ABMs) are a flexible class of computational models that simulate the dynamic behavior of individual agents. ABMs are the most widely adopted Lagrangian model, where individual people (or agents) interact and evolve in a simulation that, hopefully, resembles the true state of the population of interest [1, 2]. These models explore the theoretical interaction and evolution of individuals and have been used to simulate systems

ranging from transmissions of disease to the spread of culture. Epidemiologists have used agent-based models to track a variety of disease outbreaks, such as the propagation of the H5N1 influenza [3] and the 2014-2016 Ebola virus epidemic [4].

Since the early 1970s, researchers have noted the incredible accuracy of ABMs to model systems in which traditional modeling methods fail. In addition, the agent-based approach is attractive because ABMs provide inference at the individual level, which can be scaled up to make population-level inference.

One of the primary benefits of using ABMs to model disease is to obtain inference that can be tailored for each individual within a heterogeneous population. Many researchers have identified this benefit and have discussed how to make flexible, scalable epidemic models [5] or have argued for the widespread adoption of Lagrangian models in epidemiology [6, 7]. Some epidemic ABMs explore aspects of social interaction [8] or consider spatially explicit covariates [9, 10], and, consequently, they report Lagrangian inference that would be unobtainable through the implementation of Eulerian models.

Since ABMs have existed in theory since the 1940s, it should be no surprise that they have modeled myriad phenomenon throughout recent years. Defense experts have modeled the combat of biological warfare with these Lagrangian models [11]. City planners and environmental engineers have modeled water flow and management with ABMs [12]. Sociologists have used ABMs to analyze armed conflict and population change [13].

Due to the incredible flexibility of agent-based models, researchers in different disciplines face unique challenges and objectives in working with ABMs. Since ecological data sets are often published at the individual-level, ABMs seem to be a suitable class of models to obtain ecological inference. Usually, ecologists hope to make Eulerian inference, where they report information for species or an ecosystem as a whole. Therefore, they could analyze Lagrangian data to obtain Eulerian inference. Several ecologists have used this method of aggregating Lagrangian data such as when analyzing worm populations [14], bird populations [15], or the general patterns and theories of adaptive behavior [16]. Inspired by the

Lagrangian aggregation that ecologists have demonstrated, epidemiologists have also been able to make both individual- and population-level inference from ABMs [17, 18].

Agent-based models have also received much attention within spatial statistics and the environmental sciences. Several review papers have published applications of spatial ABMs to a variety of problems [19]. Geographic information systems (GIS) make heavy use of spatially explicit analyses, which have been implemented in ABMs [20]. Additionally, biostatisticians have analyzed human health through the use of spatial ABMs [21]. Spatial agent-based models are quite powerful, providing further personalized inference for an individual given their geographic location.

Although agent-based models can provide both Lagrangian and Eulerian inference, they have a terrible drawback: long run-time. Simulating from ABMs is more computationally expensive than Eulerian models. This can be attributed to an ABM's use of many parameters (or large parameter spaces), Markovian dynamics, intricate deterministic relationships, or a massive agent population. Even with the significant computational power found in many supercomputers, complex ABMs may require hours, days, or even weeks to complete a single simulation. Typically, researchers need to run thousands to hundreds of thousands of simulations to explore the entire parameter space and average over the model. Thus, modeling a phenomena that is remotely complex may be too computationally expensive for a study using an ABM.

Although agent-based models remain computationally expensive, much progress has been made in the understanding of an ABM's computational expense [22]. Efforts have been made to parallelize ABMs to aid in computational efficiency through the use of multi-core clusters [23, 24]. Other work has reported that reasonable constraints may result in highly efficient ABMs, such as rejecting simulations that do not accurately reflect reality [25]. Several researchers have argued for scalable agent-based simulations that can decrease the agent population without suffering from a loss in accuracy [26, 27]. Although the implementation of these ideas would decrease run-time for a complex ABM, the resulting

agent-oriented process will most likely still be more computationally expensive than an Eulerian model.

In the last few years, surrogate models have been applied to ABMs [28]. If a scientist were to record different sets of parameter inputs and the ABM's corresponding output, they can train a surrogate model (or emulator) to predict the ABMs output for a novel set of inputs. If well-trained, the surrogate model will contain all of the statistical knowledge that the ABM has without the large amount of moving parts. Therefore, a researcher may replace their original ABM with the trained surrogate model to obtain the rest of their simulations. For a study that requires several thousand simulations or more, the use of a surrogate model will reduce run-time greatly with the small cost of additional approximation. One drawback from using a surrogate model is that the modeler may not understand how the surrogate model's components work together to obtain the inference, which may resemble the notorious black-box problem [29].

In addition to computationally expensive simulation, ABMs have another rather severe drawback. An accessible and implementable regularization theory for ABMs does not exist, unlike nearly every other class of models within statistics. The lack of such a theory that is rather ubiquitous within scientific inference indicates a dearth of attention towards model selection procedures for ABMs. Therefore, researchers may not obtain insightful metrics that signal which ABM provides better predictive accuracy over another.

Despite the inability to regularize ABMs and quickly simulate a complex, realistic system, agent-based models are a valuable tool to model phenomena, such as epidemics, at the individual-level.

3 The Ecological Fallacy

Consider a heterogeneous global population that is experiencing an incredibly infectious disease. The cumulative confirmed cases are provided, as well as an infection probability.

Perhaps the infection probability is around 48%, and that number circulates the global news network. Billions of people believe in that number, thinking that they have a 48% chance to become infected. However, that probability applies to only the averaged global citizen: a right-handed, Han Chinese male that is 28 years old [30].

The idea that 48% applies to everyone is outlandish, yet it is a perfect example of the ecological fallacy. A few of the following factors can significantly affect one's infection probability: age, medical history, gender, behavior, location, environment, sanitation, and more. Even one of these factors can heavily modify that 48%. In an attempt to address any doubt on the previous claims —and in an attempt to further motivate the adoption of a Lagrangian perspective, a simulated example is provided that reveals the potential effect one's age could play on their infection probability.

3.1 A Simulation

Let n be the size of some studied population and $i(= 1, ..., n)$ denote the ith individual in this population. Suppose that five characteristics are recorded for each individual: age, exposure to others, quality of health, sanitary behavior, and gender. For this simulation, suppose that the age of individual i is obtained with

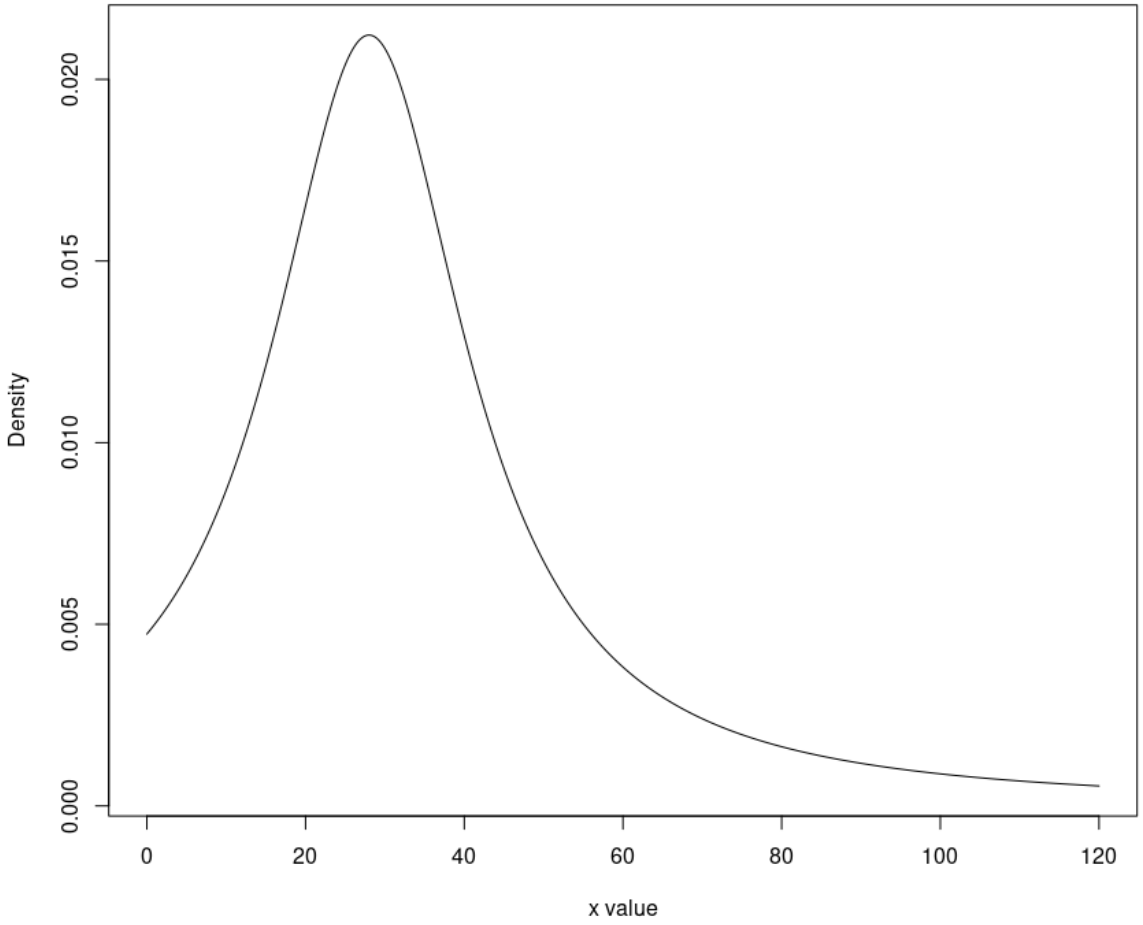
$$a_i \sim \text{Cauchy}(l, s),$$

where l and s denote the location and scale parameters of the Cauchy distribution, respectively. The Cauchy distribution is selected because the distribution of ages in a heterogeneous population are likely to be unimodal with relatively thick tails at both ends of the spectrum. In this simulation, there are various reasonable selections for l and s; for the sake of reproducibility, let $l = 28$ and $s = 15$. The density of the corresponding Cauchy distribution is depicted in Fig. 1.

The characteristic of exposure to humans, e_i, is collected because the more an individual

Figure 1: Cauchy Density

Density of the Cauchy distribution with location parameter $l = 28$ and scale parameter $s = 15$.



is interacting with potential carriers of an infectious disease, the more likely they are to contract the disease. Since this characteristic can be uniform throughout a heterogeneous population, it may be reasonable to estimate the exposure rate for individual i as

$$e_i \sim \text{Uniform}(0, 1)$$

The use of the uniform density here implies that any level of exposure is equally probable.

As most would expect, the quality of one's health, h_i, is somewhat correlated with their age, so assume health quality is a function of age a_i. Suppose that newborns and the elderly have the poorest quality of health. Further, assume that one's quality of health increases drastically between birth and teenage years, and it tapers off slowly as one ages.

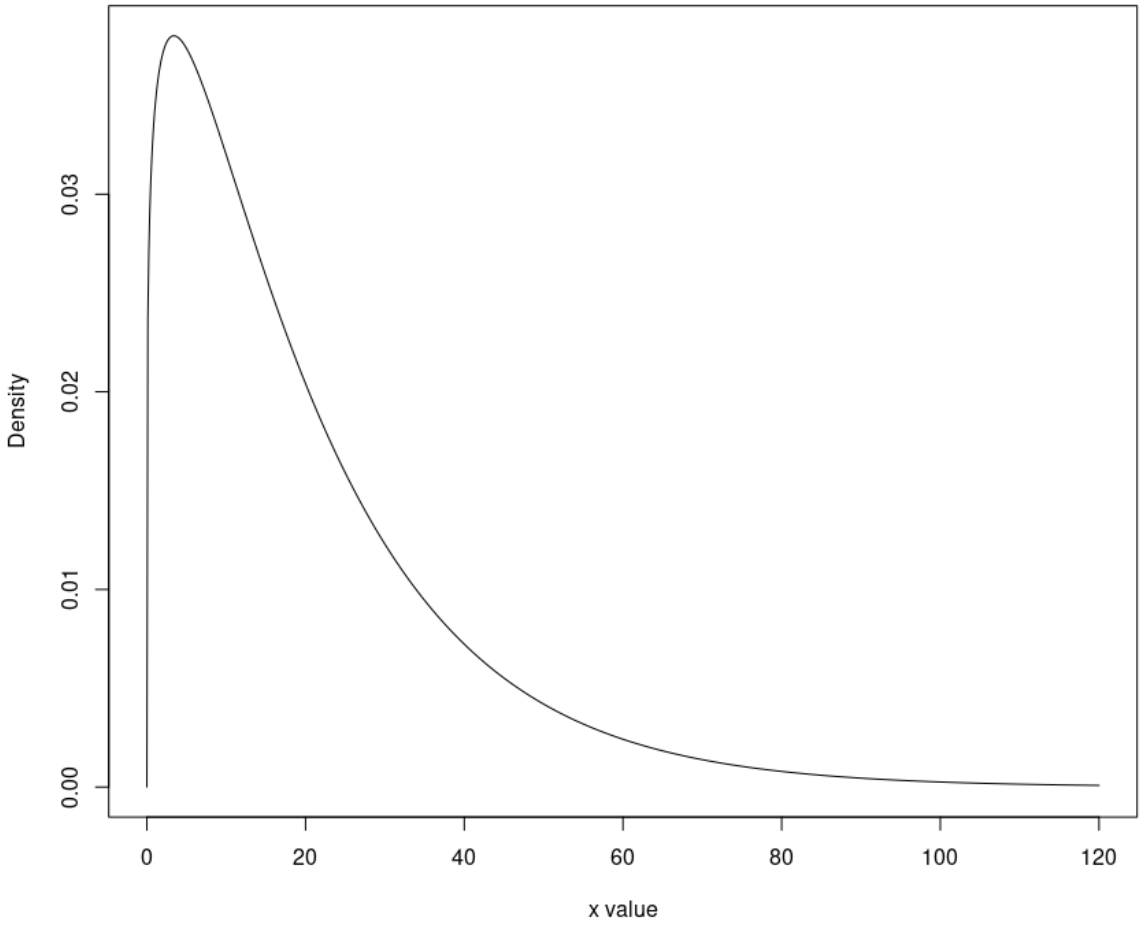
A reasonable function to determine an individual's quality of health may be provided by the density function of the Gamma distribution. We assume

$$h_i = \frac{\beta^\alpha}{\Gamma(\alpha)} a_i^{\alpha-1} e^{-\beta a_i},$$

where α and β denote the shape and rate of the Gamma distribution, respectively. While these parameters can take on many reasonable values, they are set at $\alpha = 1.2$ and $\beta = \frac{1}{17}$ for this simulation. The density of the corresponding Gamma distribution is depicted in Fig. 2.

Figure 2: Gamma Density

Density of the Gamma distribution when the shape parameter $\alpha = 1.2$ and the rate parameter $\beta = \frac{1}{17}$. Note that the density is highest near lower values of x. Therefore, for lower ages, h is higher, implying that the youth have a higher quality of health on average than the elderly in this simulated population.



Similar to exposure to others, sanitary behavior, s_i, is a characteristic that is independent of age. Perhaps there are some individuals with poor sanitation and some with excellent sanitation, but the majority are adequate. Suppose sanitary behavior can be modeled with

$$s_i \sim \text{TruncNorm}(a, b, \mu, \sigma),$$

where μ and σ are the mean and standard deviation respectively of the normal distribution, and a and b are bounds that truncate the simulations. Note that a normal distribution or a uniform distribution are sensible choices, as well. For the sake of this simulation, let $a = 0$, $b = 1$, $\mu = 0.5$, and $\sigma = 0.15$. Since this truncated normal distribution is symmetric, without loss of generality, assume that the greater the value of s_i, the worse an individual's sanitary behavior. The density of the corresponding Truncated Normal distribution is depicted in Fig. 3.

The last characteristic to be simulated is gender. Suppose individual i's gender, g_i, is independent of all previous characteristics and is obtained with a simple Bernoulli draw with equal probability of being male or female:

$$g_i \sim \text{Bernoulli}(0.5)$$

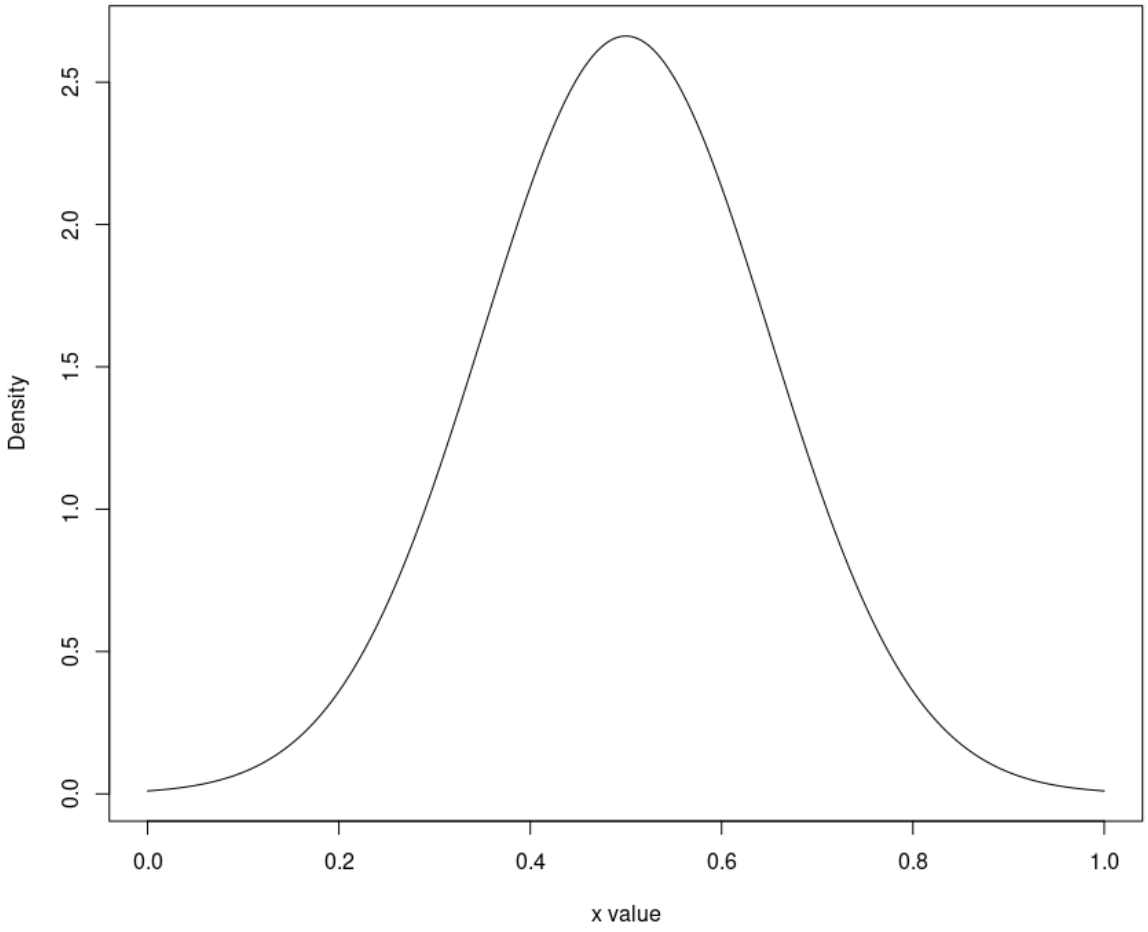
Suppose that males have a slightly greater chance of becoming infected than females. If $g_i = 1$ denotes that individual i is male, gender is positively related to infection probability.

Several of the aforementioned characteristics lie in the interval $[0, 1]$, which is convenient if each individual's infection probability is obtained as a weighted sum of the characteristics. Note that $e_i, s_i, g_i \in [0, 1]$. If a weighted sum is used to obtain one's infection probability, it would be pragmatic to normalize the remaining characteristics (a_i and h_i) such that each value belongs in $[0, 1]$. For age, simply take

$$a_i^* = \frac{a_{i,m} - a_{m,min}}{a_{m,max} - a_{m,min}},$$

Figure 3: Truncated Normal Density

Density of the Truncated Normal distribution with boundaries $a = 0$ and $b = 1$, mean $\mu = 0.5$, and standard deviation $\sigma = 0.15$. Note that the majority of individuals within the simulated population have adequate sanitary behavior, while few individuals have poor or exceptional sanitary behavior.



where $a_{m,min}$ and $a_{m,max}$ are the minimum and maximum value of all individuals' modified ages $a_{i,m}$:

$$a_{i,m} = \exp\left[m \cdot a_i - 1\right],$$

where m is the modulating scale for an individual's age. Since m is not biologically interpretable, $m = 3$ is arbitrarily chosen to emphasize older individuals during the scaling process. If more heterogeneity is desired in the simulation, m can be randomly sampled from a range of positive real numbers for each individual i. Note here that the induced exponential transformation is used so that as one's age a_i increases, the value of $a_{i,m}$

increases exponentially. While the normalization is necessary in the described approach, the deterministic relationship between $a_{i,m}$ and a_i is not. However, since infectious diseases appear to affect the elderly more severely than the youth, this transformation will result in exponentially larger weights toward one's infection probability.

By definition, a_i^*, e_i, s_i, and g_i are all positively related to infection probability. So, it would make sense to ensure that quality of health has a similar relationship. The current definition of quality of health is an increasing function of age, so older individuals have a higher quality of health value. Suppose this interpretation is flipped with the following transformation:

$$h_i^* = 1 - \frac{h_i - h_{min}}{h_{max} - h_{min}},$$

where h_{min} and h_{max} denote the minimum and maximum quality of health value across all individuals, respectively.

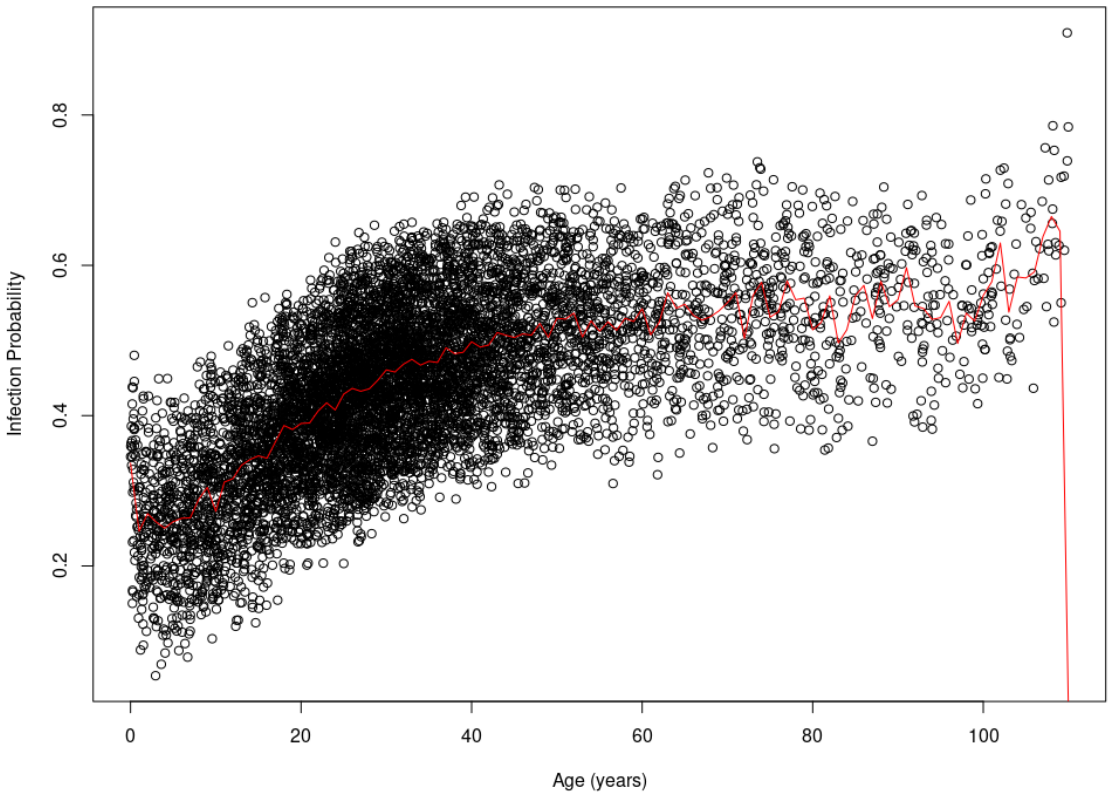
Now that all five characteristics are normalized between $[0, 1]$ and share a positive relationship with one's infection probability, the infection probability, ϕ_i, can be obtained as a weighted sum:

$$\phi_i = 0.2a_i^* + 0.2e_i + 0.3h_i^* + 0.2s_i + 0.1g_i$$

Since this is a fictional infectious disease, somewhat arbitrary weights are assigned for each characteristic. Notice that age is both a direct and indirect factor in the infection probability, since quality of health is a function of age. Fig. 4 shows how age affects one's infection probability based on 7,851 simulated individuals. Each black circle in this plot represents an individual. Since age is plotted against infection probability, one can see the general relationship. The red line follows the average infection probability as age increases. Excluding the short period following birth, individual infection probabilities increase with age.

Figure 4: Infection Probability v. Age

Infection probabilities based on 7,851 simulated individuals. Each black circle represents an individual within the simulated population. As age increases, the infection probability generally increases. The red line follows the average infection probability. Note that there are less individuals after the age of 60, so the average infection probability is more susceptible to outliers and, hence, volatile.



This disease has an average infection rate of 48%. However, a healthy 9-year old female with moderate exposure to humans and average sanitary behavior has a 28% chance of infection. An unhealthy 71-year old male with high exposure to humans and poor sanitary behavior has a 70% chance of infection. And a healthy 1-year old baby girl with excellent sanitation and low exposure to humans has a low 9% chance of infection.

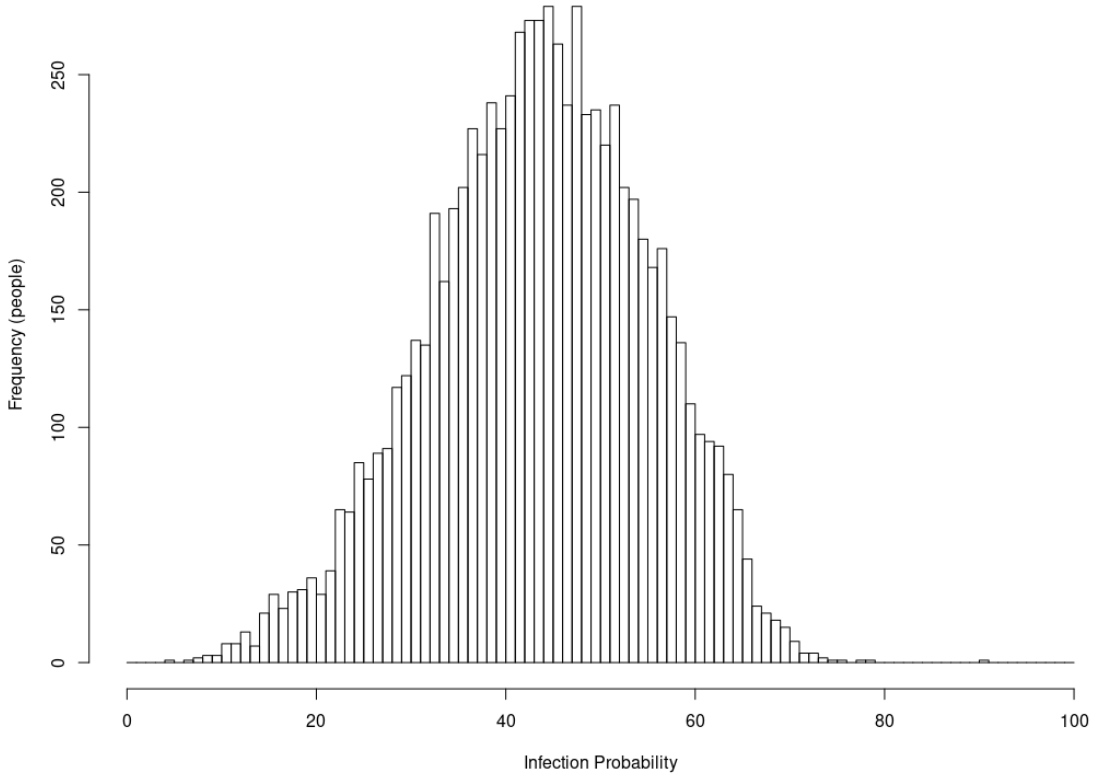
Of 7,851 simulated individuals, only 279 people (approximately 4%) have roughly a 48% chance of infection. The projected infection probability of 48% is inaccurate for the remaining approximately 96% of the population. A histogram of infection probabilities depicts the ecological fallacy in Fig. 5; only one of its bars corresponds to a 48% infection probability. The remaining bars contain the stories of the remaining 96% of the population

that lack a personalized infection probability if one were to use the 48% figure overall. In fact, it is the lack of personalization that is the major flaw in an Eulerian perspective. Any common Eulerian epidemiological model would suffer from this ecological fallacy.

Perhaps the importance of this concept can be best exemplified by the aforementioned 71-year old man. Maybe, he accounts for a 48% chance of infection when considering grocery shopping or leisure activities. However, his true infection probability is 0.70, which is significantly higher. Thus, his measurement of risk is invalid, which may result in him unnecessarily contracting the disease.

Figure 5: Infection Probability v. Frequency of People

Each bar within this histogram represents one percent of infection probability. It appears that the average infection probability lies at the unimodal peak of roughly 48%.



4 Methodology

A topic that is of particular interest during this thesis' time of completion is the SARS-CoV-2 pandemic. In what follows, an ABM and surrogate model are delineated that can be used to analyze this pandemic. These models were first proposed in one of my recent publications with Dr. Hooten and Dr. Wikle [28]. The following ABM will provide personalized inference for the passengers of the Diamond Princess cruise ship that suffered from the most disruptive pandemic in recorded history well before it reached the Western world. The surrogate model contains the statistical knowledge of the ABM and will complete the remaining desired simulations much quicker than the ABM.

4.1 Agent-based Modeling

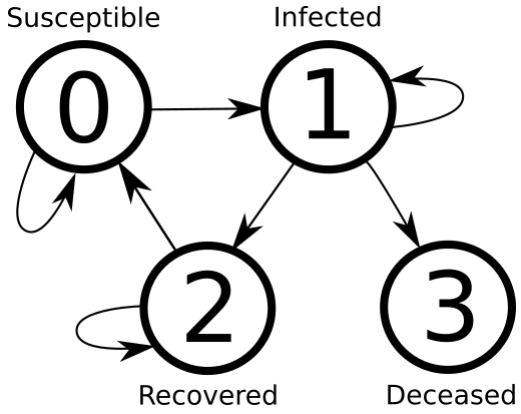
Due to the flexibility of ABMs and their agent-oriented process, there is not a direct way to document the exact model in some concise statistical notation. However, a description of the statistical and deterministic relationships follow.

Let $z_{i,t} \in \{0, 1, 2, 3\}$ correspond to the state of individual $i(= 1, ..., N)$ at time $t(= 1, ..., T)$, where state 0 is 'susceptible', state 1 is 'infected', state 2 is 'recovered', and state 3 is 'deceased'. Define $\pi_{j,k,i,t} \equiv P(z_{i,t} = k | z_{i,t-1} = j)$. Let $\pi_{0,1,i,t} = \phi_{0,1,i,t}$ denote the probability that individual i, who is susceptible at time $t - 1$, becomes infected at time t. Similarly, let $\pi_{0,0,i,t} = 1 - \phi_{0,1,i,t}$ be the probability that individual i, who is susceptible at time $t - 1$, remains susceptible at time t. Assume that an individual cannot transition between two states in a given day, so set $\pi_{0,2,i,t} = \pi_{0,3,i,t} = 0$. Additionally, assume that an individual cannot reverse to a previous state without having recovered: $\pi_{1,0,i,t} = 0$. Since death is an absorbing state, set $\pi_{3,3,i,t} = 1$ and $\pi_{3,0,i,t} = \pi_{3,1,i,t} = \pi_{3,2,i,t} = 0$. A diagram that depicts the possible state transitions can be found in Fig. 6.

Lastly, $\pi_{1,3,i,t} = \phi_{1,3}$ denotes the probability of infection-induced death for each unit of time. Note that this probability does not depend on the individual nor time. However, the

Figure 6: Susceptible-Infected-Recovered-Deceased Transitions

Susceptible individuals can either remain susceptible or become infected. Infected individuals can remain infected, recover from infection, or become deceased. Recovered individuals may remain in the recovery state or become susceptible again. Deceased individuals remain deceased; death is an absorbing state.



cumulative probability that individual i becomes deceased due to infection increases for each unit of time that they are infected. Let τ_i be the maximum number of units of time that individual i remains infected.

As stated in the literature review, the true power of ABMs are their ability to account for semi-Markov or non-Markov dynamics. This ABM relies on a semi-Markov dynamic for recovery, where an individual i will recover in τ_i days if they do not die during infection. Since residence times in the infected state likely vary between individuals, suppose τ_i is stochastic: $\tau_i \sim Pois(\lambda)$, where λ is the population-wide intensity associated with recovery time. Therefore, the cumulative probability that an individual i dies from the infection varies, whereas the daily probability of death is constant for all agents.

Like most epidemic studies, there is a primary interest in $\phi_{0,1,i,t}$ and $\phi_{1,3}$. There may also be interest in estimating the probability that individual i recovers from infection on day t: $\pi_{1,2,i,t}$. Clearly, the defined transition probabilities are Markovian and imply geometric residence time in each state.

Let the disease spread among individuals by inducing dependence in the transition from 'susceptible' to 'infected' with probability

$$\phi_{0,1,i,t} = \text{logit}^{-1}\left[\frac{\beta_0}{N}\left(\sum_{i\neq j} I_{\{z_{j,t-1}=1\}} - N_\nu\right)^2 + \beta_1\right], \tag{1}$$

where $\beta_0 < 0$ and $I_{\{z_{j,t-1}=1\}}$ is an indicator equal to 1 when the j-th individual at time $t-1$ is infected and 0 otherwise. Therefore, the probability of an individual becoming infected is dependent on the number of infected individuals on the previous day. Additionally, the infection probability function will rise until its peak when the number of infected individuals reaches N_ν. As a constant, N_ν coincides with a hypothesized inflection point in the cumulative number of confirmed cases for the population of interest. After this point, the infection probability function will begin to decline. The value of N_ν can correspond with the behavior of agents, as well as any relevant policies or stay-at-home orders. This unimodal function should approximate the relative infection probability's rise until agents were quarantined. The peak of $\phi_{0,1,i,t}$ should coincide with the inflection point in the epidemic. Lastly, parameters β_0 and β_1 have no scientific interpretation in this model; they merely serve as values that control the simulated spread of SARS-CoV-2.

To simulate the spread of a novel pathogen, set $z_{i,1} = 1$ for $i = 1, ..., 5$ individuals. Then, set $z_{i,1} = 0$ for $i = 6, ..., N$, where N is the total population. Since most data sources for SARS-CoV-2 report population-level data, all output of this spatial ABM will be aggregated to the population level. That is, the cumulative new infected individuals will be obtained on day t as $n_t = \sum_i \sum_{t'=1}^{t} I_{\{z_{i,t'-1}=0, z_{i,t'}=1\}}$. The simulated output for n_t should be comparable to the data gathered by reputable sources, such as Japan's Ministry of Health, Labour, and Welfare or the Southern Nevada Health District. Thus, a comparison of n_t at each day and the widely-available SARS-CoV-2 data dashboards will reveal whether this ABM is simulating outbreaks relatively well.

4.2 Surrogate Modeling

Since the aforementioned agent-based model is rather computationally expensive, a surrogate model is used to interpolate inference for simulation parameters that have not been tested. To begin, 1,000 simulations are run on the ABM. Most Bayesian studies prefer samples of 20,000 or more simulations. Thus, the remainder of the simulations will be obtained through the computationally efficient surrogate model.

A surrogate model is an auxiliary model that follows the ABM; the set of recorded inputs and outputs from the ABM is used to estimate the surrogate model's parameters. A well-constructed surrogate model contains all of the statistical information that the ABM will contain. Additionally, the surrogate model will be able to complete simulations much quicker than the ABM.

Let y_t represent the observed cumulative new infections on day $t(= 1, ..., T)$ and denote $\mathbf{y} = (y_1, ..., y_T)'$ as the observed data. If this method works well, $\mathbf{y}$ (the observed data) and $\mathbf{n} = (n_1, ..., n_T)'$ (the simulated ABM output of cumulative confirmed cases) should be nearly identical. Consider a Gaussian process for the surrogate model

$$\mathbf{y} \sim N(\mathbf{K}\tilde{\mathbf{n}}, \sigma^2\mathbf{I}), \tag{2}$$

where $\tilde{\mathbf{n}} = (\tilde{n}_1, ..., \tilde{n}_T)'$ denotes the cumulative new cases for all days $t \in \{1, ..., T\}$ and $\mathbf{K}$ denotes an $n \times T$ mapping matrix. Surrogate models fall under two main categories: first-order and second-order emulators. A second-order emulator contains a covariance term, whereas a first-order emulator does not. The above Gaussian process is best categorized as a first-order emulator. Although there is no covariance modeled in $\mathbf{y}$, covariance is modeled in $\tilde{\mathbf{n}}$.

For some set of input $\boldsymbol{\theta}^{(l)}$, the ABM is used to simulate output $n_t^{(l)}$ for $l = 1, ..., L$. Then, the simulation's output is aggregated into a vector $\mathbf{n}^{(l)} = (n_1^{(l)}, ..., n_T^{(l)})'$. Assume that $\tilde{\mathbf{n}}$ can be characterized by another Gaussian process with the distribution $\tilde{\mathbf{n}} \sim N(\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}), \tilde{\boldsymbol{\Sigma}})$

described by

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \sum_{l=1}^{L} w^{(l)}(\boldsymbol{\theta})\mathbf{n}^{(l)}, \tag{3}$$

$$\tilde{\boldsymbol{\Sigma}} = \sigma_n^2 \exp\left(-\frac{\mathbf{D}_n}{\gamma_n}\right), \tag{4}$$

where $\exp(\mathbf{D})$ takes the exponent of each element in the matrix $\mathbf{D}$ and $\boldsymbol{\theta}$ is a vector that contains all parameters in the ABM that are unknown. Since the conditional mean of the latent process $\tilde{\mathbf{n}}$ is a weighted average of the analogues $\mathbf{n}^{(l)}$, this type of surrogate model is referred to as an 'analogue emulator.' The weights can be set as a function of proximity between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(l)}$ in parameter space

$$w^{(l)} = \frac{\exp\left(-(\boldsymbol{\theta} - \boldsymbol{\theta}^{(l)})'\boldsymbol{\Gamma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(l)})\right)}{\sum_{l=1}^{L} \exp\left(-(\boldsymbol{\theta} - \boldsymbol{\theta}^{(l)})'\boldsymbol{\Gamma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(l)})\right)}, \tag{5}$$

where $\boldsymbol{\Gamma}_\theta$ is the modulating range parameter.

Note that the temporal covariance matrix (4) depends on pairwise temporal differences in the $T \times T$ matrix $\mathbf{D}_n$. Additionally, it accounts for dependence in the process $\tilde{\mathbf{n}}$ not accounted for by the analogues.

A surrogate model framework has two stages. In the first stage, $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\Sigma}}$ are estimated by calibrating the surrogate model using the ABM simulation's input and output. To calibrate this emulator, an aggregated loss function is optimized with respect to γ_n and $\boldsymbol{\Gamma}_\theta$. Then, the product of the emulator density functions is maximized over all $\mathbf{n}^{(l)}$ while conditioning on $\{\boldsymbol{\theta}^{(-l)}\}$ and $\{\mathbf{n}^{(-l)}\}$, which is the set of parameters and analogues without the l-th simulation. Thus, each analogue $\mathbf{n}^{(l)}$ may be interpreted as data dependent on the other analogues to aid the learning on the smoothness in the distribution of the analogues.

Minimal uncertainty is assumed in γ_n and $\boldsymbol{\Gamma}_\theta$ since the computer experiment can be made as significantly large. Thus, they are treated as constants when fitting the surrogate

model (2) using an MCMC algorithm. In the second stage of this surrogate implementation, the ABM parameter vector $\boldsymbol{\theta}$ is updated using Metropolis-Hastings sampling. Lastly, Gibbs updates are used for $\tilde{\mathbf{n}}$ and σ^2.

5 The Diamond Princess Cruise Ship

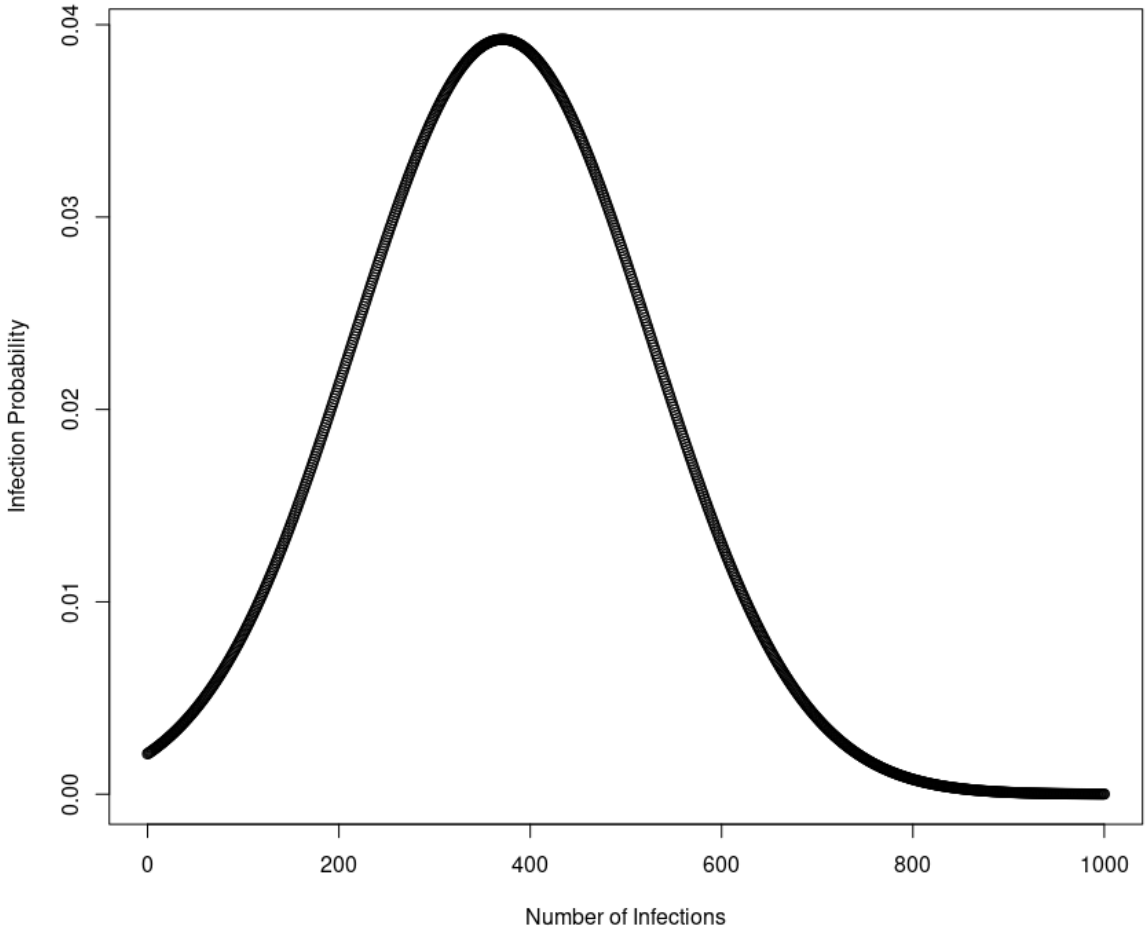
Although some of the Western world was aware of the ongoing SARS-CoV-2 pandemic in China, most did not initially concern themselves with the virus. However, the Diamond Princess garnered global attention as SARS-CoV-2 rapidly spread from cabin to cabin. Within one week of quarantine off the coast of Japan, the Diamond Princess accounted for the majority of reported cases outside of mainland China [31]. This case study was first presented in a previous paper on which I was a co-author [28].

On January 20, 2020, the Diamond Princess departed from the Port of Yokohama in Japan with passengers from all over Asia. One such passenger from Hong Kong was believed to be the initial case on the cruise ship. By February 1st, the ship was quarantined at Naha Port in Okinawa. With limited resources to test passengers, cases were not confirmed on the Diamond Princess until February 5th, when there were already ten confirmed cases. One day later, there were 20 confirmed cases. On February 7th, there were 61 cases. SARS-CoV-2 thrived among the tight quarters of the cruise ship and had alerted the entire world of its presence.

Although the Diamond Princess was not immediately quarantined after the first infection, the cruise ship had a nearly closed population since it departed from the Port of Yokohama. For this reason, the cruise ship makes an excellent case study for the ABM-surrogate model constructed in the previous section. Additionally, the delayed quarantine provides an opportunity to test the parabolic nature of the infection probability, since the probability likely increased until a peak and then decreased. Presumably, the peak coincides with the implementation of a quarantine or shortly after, since the infected individuals may take a

few days to register as confirmed cases. The infection probability from equation (1) is visualized in Fig. 7.

Figure 7: Infection Probability v. Number of Infections aboard the Diamond Princess
After optimizing for $\boldsymbol{\beta}$ in equation (1), we let the number of infections $\sum_{i \neq j} I_{\{z_{j,t-1}=1\}}$ vary. As the number of infections increases between 0 and 380, an individual's infection probability ($\phi_{0,1,i,t}$) increases. However at 380 infected individuals, the infection probability peaks. This coincides with the number of infections shortly after the Diamond Princess quarantine. Following the quarantine, infection probability decreased.



While the confirmed cases on the cruise ship are readily available in a compact table on Wikipedia, the data was gathered directly from Japan's Ministry of Health, Labour, and Welfare [32–45]. The data contained cumulative confirmed cases and cumulative tests administered. The last available report of confirmed cases aboard the Diamond Princess came out on March 5th —over two weeks after the previous report. However, some of the new confirmed cases may have been contracted after passengers with negative test results

left the ship; this process began on February 19th. Thus, the last report was excluded from this case study. Additionally, the Diamond Princess staff failed to report testing results for several dates following the quarantine.

Minimal data preparation was needed for the ABM. Dates were assigned to the cumulative confirmed cases, and T was obtained as the length between the date of the first report and the last report. The Ministry's reports declared there were roughly $N = 3,711$ people on the ship throughout the quarantine with minor changes.

Although N and T are directly measured from the compiled data set, λ, $\phi_{1,3}$, and $\boldsymbol{\beta}$ still need to be declared for the ABM. Since λ denotes the average time of recovery for an infected individual, a reasonable choice would be $\lambda = 13.5$ [46]. For the remainder of this manuscript, assume $\lambda = 14$ to induce more variation in recovery time. Additionally, since the Diamond Princess had an older population than that studied in [46], it is likely that the recovery time for the average passenger exceeded that of an average global citizen.

While the probability of mortality varies from person to person, the average probability of death for a passenger would be around 3% [47]. Since death is geometric, $\phi_{1,3}$ is the probability such that $P(x \leq 14) = 0.03$ when $P(X \leq k) = 1 - (1 - \phi_{1,3})^k$ for discrete values k. From this, $\phi_{1,3} = 0.002$ is estimated as the average passenger's daily probability of death. Recall the simulated study in Section 3, where the probability of infection varies between people. A similar argument can be made regarding the probability of death. However, the event of death, $d_{i,t}$, for each agent in the ABM is determined each day from a Bernoulli draw:

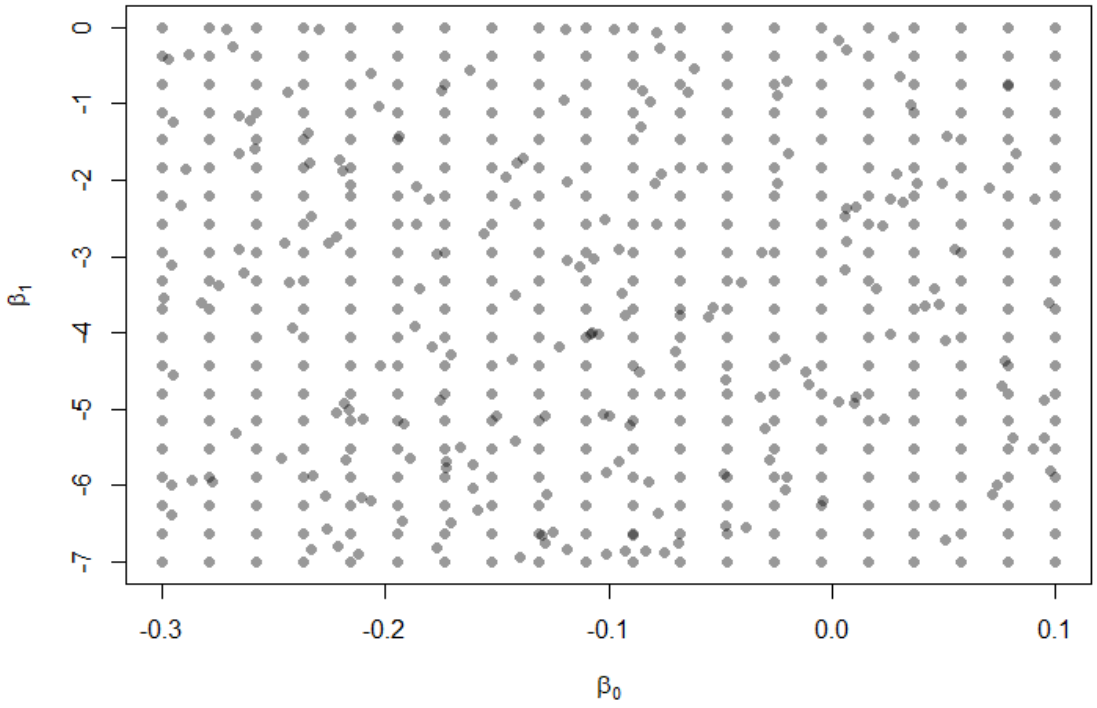
$$d_{i,t} \sim \text{Bernoulli}(\phi_{1,3})$$

Therefore, the probability of death for each agent on each day is stochastic rather than deterministic. This semi-Markovian dynamic not only accurately reflects the true state of the passengers, but it also demonstrates a convenient dynamic that can rarely be implemented in an Eulerian model.

Both λ and $\phi_{1,3}$ are biologically interpretable from global data. That is, λ is not expected to greatly vary from one heterogeneous population to the next, and a similar argument can be made about $\phi_{1,3}$. However, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ tunes the infection probability, which is likely to change between any population due to societal behavior, population density, and many other factors. There is not a single reasonable proposal for $\boldsymbol{\beta}$. Rather, an entire parameter space is proposed to cover all potential values of $\boldsymbol{\beta}$. Within this parameter space, there are 400 equally spaced sets of $\boldsymbol{\beta}$ and 400 randomly selected sets of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ within $(-0.3, 0.1) \times (-7, 0)$. Fig. 8 displays the different sets of $\boldsymbol{\beta}$ used in the ABM simulations.

Figure 8: $\boldsymbol{\beta}$ Parameter Space

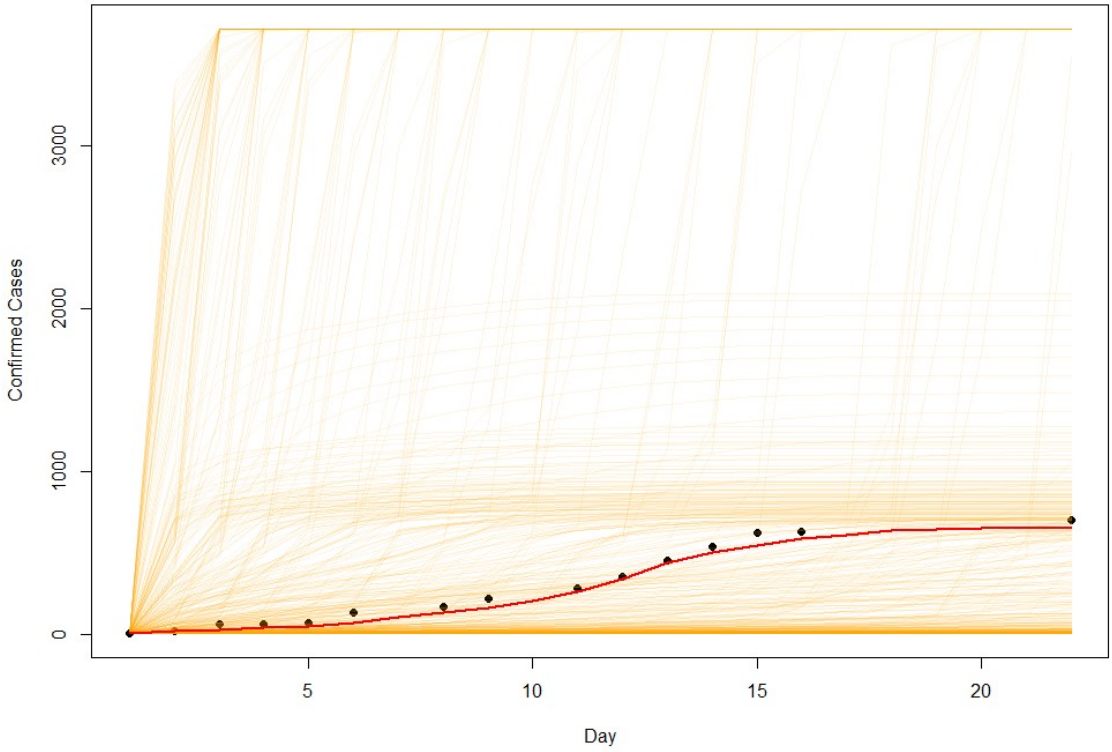
Each dot in the figure corresponds to a pair (β_0, β_1). All of these pairs of $\boldsymbol{\beta}$ were used in the initial ABM simulations. Notice that there is a structured grid of points throughout the depicted parameter space of (β_0, β_1), as well as random pairs of $\boldsymbol{\beta}$ throughout. The combination of this structured grid and random points contributes to an adequate exploration of the parameter space.



With these 800 different sets of $\boldsymbol{\beta}$, the ABM obtains 800 simulations with unique infection probabilities. The output of these simulations is visualized in Fig. 9. Each faint orange line is the output of one simulation, and the black points are the actual cumulative cases aboard the Diamond Princess. The thick red line is the averaged simulation across

all 800 ABM simulations. Note how close the average simulation is to the actual data points for the Diamond Princess. This signals that the ABM captures the dynamics between passengers on the Diamond Princess incredibly well.

Figure 9: ABM Simulations for Confirmed Cases on the Diamond Princess
Each thin orange line tracks the number of confirmed cases as the days increase. The black points are the actual data that was reported by the Diamond Princess cruise ship staff. The thick red line that runs through these black data points represents the average ABM simulation.



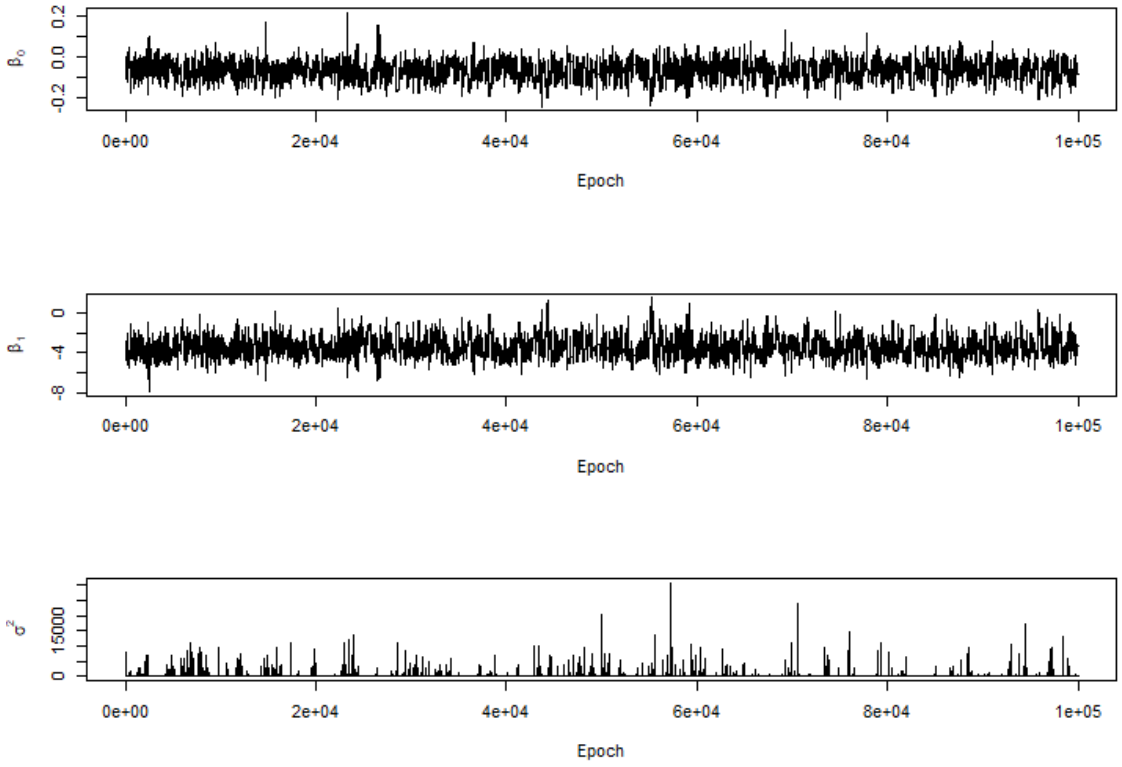
Once the ABM's simulated output is obtained, the surrogate model is given the 800 sets of $\boldsymbol{\beta}$ with their corresponding ABM output. Now, the objective is to optimize the emulator's parameters: γ_n, σ_n^2, and $\boldsymbol{\Gamma}_\theta$. This is done in the R programming language by minimizing the negative log-likelihood of the emulator [48]. Once the parameters are optimized, the surrogate model's predictions can be tested by simply providing some temporary values for $\boldsymbol{\beta}$.

At this point, the surrogate model is tasked with quickly completing the remaining simulations using the information obtained from the ABM, which is done using an MCMC algorithm. The only new information that must be introduced at this stage are the tuning

values for $\boldsymbol{\beta}$, since a random-walk approach is utilized. A total of 100,000 simulations are completed with the MCMC algorithm, which uses a Metropolis-Hastings method for $\boldsymbol{\beta}$ proposals and conjugate updates for σ^2 and the cumulative confirmed cases $\tilde{\boldsymbol{n}}$ in (2). Trace plots for $\boldsymbol{\beta}$ and σ^2 are provided in Fig. 10, which reveal both good mixing and convergence as the epoch (or, number of iterations) increases.

Figure 10: Trace Plots for $\boldsymbol{\beta}$ and σ^2

These trace plots track the values of β_0, β_1, and σ^2 as the number of MCMC iterations increase. The relatively low volatility in these trace plots indicate that there is both good mixing and convergence in the MCMC algorithm.



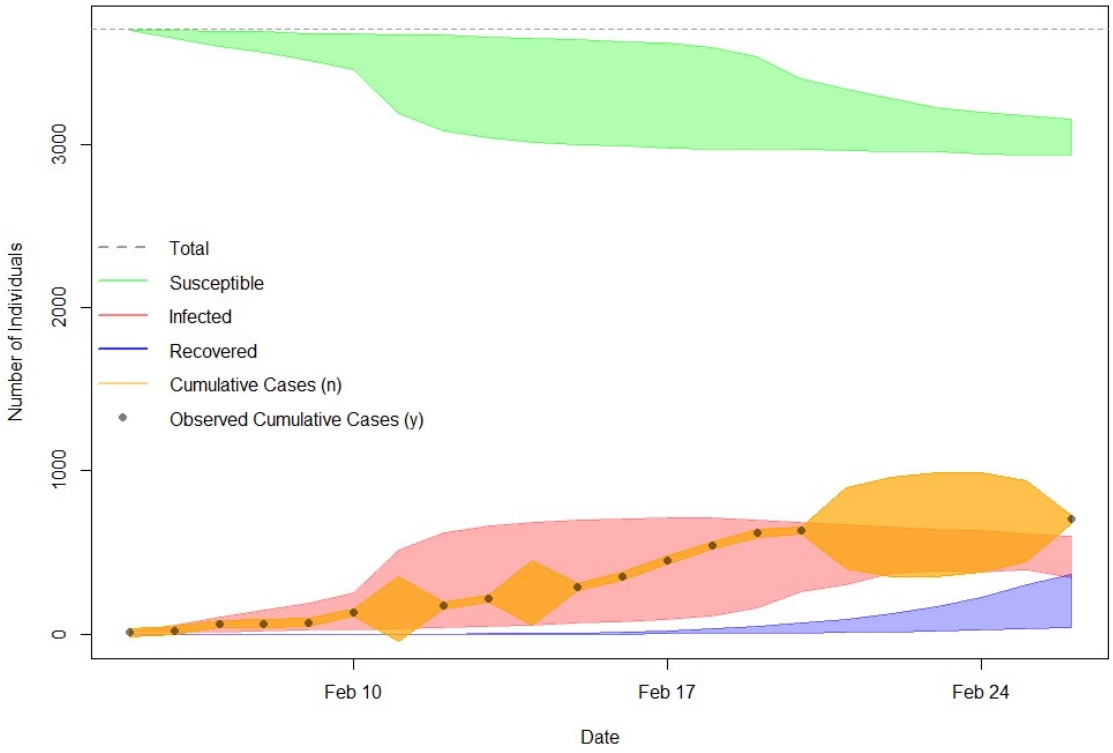
The final results are aggregated into Fig. 11, which displays the following population compartments: susceptible (green), infected (red), recovered (light blue), and cumulative confirmed cases (orange). The polygon shapes encapsulate the 95% confidence region belonging to each compartment throughout the study period. The observed Diamond Princess data is represented with black dots. Note that the cumulative cases polygon perfectly follows the actual data. This is quite incredible considering the surrogate model was only given 800 sets of input and output from the ABM, which was never given the

actual data.

The surrogate model had to interpolate for days in which results were not reported by the Diamond Princess staff. Notably, February 11th, 14th, and 21st-25th lacked testing reports. This ability to interpolate within the testing dates is an attractive feature that this Lagrangian model shares with many Eulerian models.

Figure 11: Population Compartments for the Diamond Princess

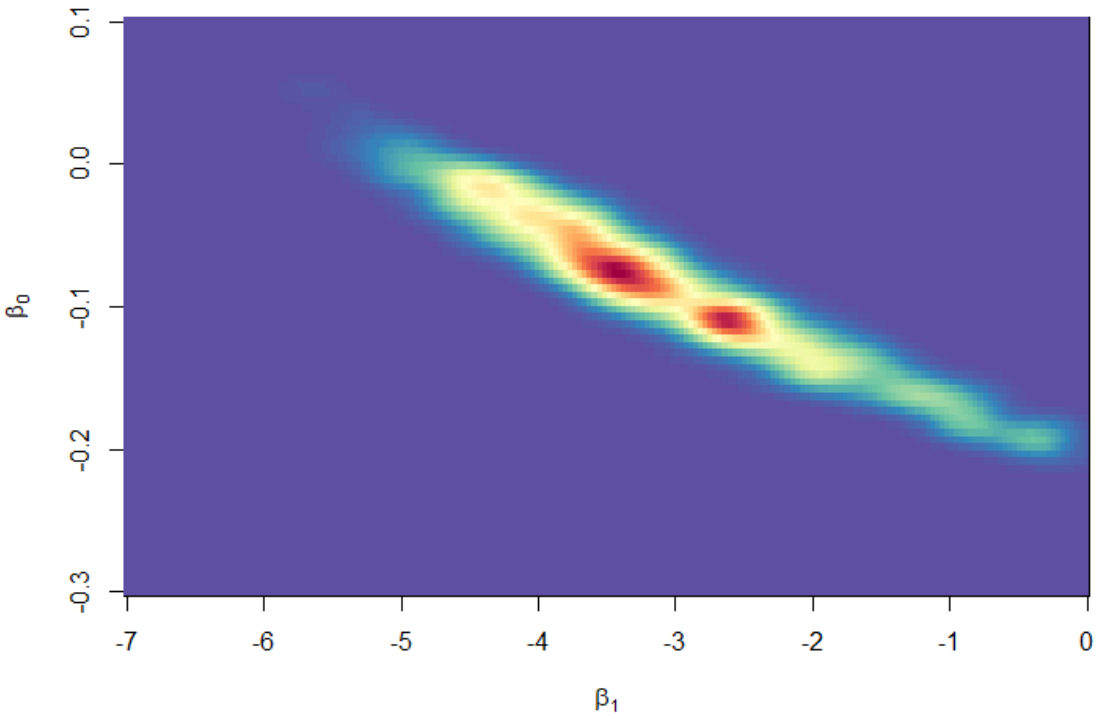
The green region represents the 95% confidence region for the number of individuals that are susceptible at any date throughout the study. The red region represents the 95% confidence region of infected individuals, the orange region represents the 95% confidence region of cumulative cases, and the blue region represents the 95% confidence region for the number of recovered individuals.



One common objective in modeling infectious diseases and epidemics is to determine the infection probability. The infection probability function for $\phi_{0,1,i,t}$ requires scaling parameters $\boldsymbol{\beta}$. It may be of great interest to research the interpretability of these parameters; however, such an analysis extends beyond the scope of this thesis. Yet, a rather informative graphic can be shared that may aid in such pursuits. In Fig. 12, a heat map of the parameter space of $\boldsymbol{\beta}$ is documented, where dark red regions are favored by the surrogate model,

yellow regions were well-explored, light blue regions received a high rejection rate in the Metropolis-Hastings ratio, and the surrounding purple region was never explored.

Figure 12: β Heat Map for the Diamond Princess Cruise Ship
This heat map displays the frequency of acceptance in the MCMC algorithm for each pair $\boldsymbol{\beta}$. Dark red regions of the parameter space were well-accepted by the MCMC algorithm, indicating that these regions are likely to be the true values of $\boldsymbol{\beta}$. As the red transitions to purple, the regions experience less acceptance in the MCMC algorithm.



6 Scalability for Clark County, Nevada

The Diamond Princess outbreak makes for a convenient case study [28]. The ship had a closed population of no more than 4,000 people. Agent-based models become increasingly expensive as the agent population grows, so a relatively small population of 4,000 agents is quite convenient. Naturally, one would question the scalability of ABMs for larger data sets. An intuitive, and entirely correct, assumption would be that ABMs are not the most scalable class of models. To demonstrate this, Table 1 records the time for completion of just one simulation on the ABM at various population sizes. These simulations were run

Table 1: Run-times for an ABM Simulation with Various Population Sizes

Population Size	Run-time (s)
200	0.03
500	0.15
1,000	0.21
5,000	3.31
10,000	12.51
15,000	27.54
25,000	75.45
50,000	301.25
100,000	1191.44
500,000	29582.14
1,000,000	112782.37
1,500,000	299683.27
2,000,000	2328723.71

on an Intel Core i7-8700 processor in the Windows 10.0.19041 environment.

Notice that the run-time for a single simulation on the ABM is polynomial. Such a run-time is inefficient when dealing with sufficiently large data sets. For most Bayesian studies, it is typical to have 20,000-50,000 or more simulations. This ABM is evidently computationally expensive and requires a significant amount of time. One possible remedy is the use of a surrogate model which has a linear run-time. However, a sizable amount of simulations must be run on the ABM so that the surrogate model will be able to inherently learn the dynamics between agents in an Eulerian fashion. If one were to restrict the number of ABM simulations to just 500 —which may not render great approximation for the surrogate model, it is likely that an agent population of anything more than 100,000 agents would require nearly a week to finish.

6.1 An Extension to Clark County

Consider the application of the ABM-surrogate model to the SARS-CoV-2 outbreak in Clark County, Nevada. The county has roughly 2,267,000 residents. This population size should seem daunting given the results in Table 1, where just 10,000 agents would take

nearly 3 hours to simulate 800 times.

Most Eulerian models would be able to scale the population down such that inference can be readily obtained. This inference can then be scaled back up to the true population size. However, such an approach does not necessarily work in a Lagrangian framework, despite sometimes needing a scaled population. If the population were not scaled in this Lagrangian model, the epidemiologist or biostatistician would receive their results likely after the pandemic has passed.

Suppose one scales the Clark County population by a factor of 0.001 to obtain an agent population of 2,267. This scaling factor must be applied to both the population and the confirmed cases. Since the confirmed Clark County cases are 5 on January 20th, 36 on March 1st, 1,017 on March 23rd, and 5,980 on May 15th, the scaled confirmed cases would be 0.005, 0.036, 1.017, and 5.98, respectively. The ABM attempts to capture the true dynamics of the pandemic, so the model strictly uses discrete data. There is never 5.98 cases of SARS-CoV-2 —just 5. By definition, the ABM rounds cases down, since it does not consider "partial" cases as a complete confirmed case. For this reason, the ABM considers just 1 infection of 2,267 people on March 23rd.

Although the scale is approximately correct, there are two fundamental problems that exist. The first problem is that Clark County first reported confirmed cases on January 20th, 2020 —the same day that the Diamond Princess embarked from the Port of Yokohama. Yet, the ABM does not sense any confirmed case until March 23rd. Therefore, the ABM misses out on over two months of the transmission dynamics. This period is crucial in understanding the SARS-CoV-2 outbreak, since it makes up a large part of the first wave of the Clark County epidemic.

The second problem is that, although the population and confirmed cases are scaled down, the length of period of infection remains unaltered. On April 1st, the Southern Nevada Health District reported 2,002 cumulative confirmed cases. With a scaling factor of 0.001, the ABM reads that there are 2 confirmed cases on April 1st. So, there are 9 days

between the first infection and the second infection. Since $\lambda = 14$, it is quite reasonable that the initial infected agent recovers before infecting anyone else in some of the ABM simulations. In these simulations, the outbreak of Clark County is simulated inaccurately with just one infection for a few days in March. Once the inference is scaled back up to the true population size, the results would display a seemingly-spontaneous outbreak of 1,000 cases on March 23rd. Then, those 1,000 infected residents will immediately recover altogether without infecting any other individuals before April 1st. Clearly, the ABM output would be faulty.

These two problems are ubiquitous in large data sets so long as the disparity between cumulative cases and population size is significant. If one were to apply a similar Lagrangian model to Clark County during a more prevalent pandemic, the disparity between population size and confirmed cases would be less. If so, both of the aforementioned problems would be alleviated. However, in the analysis of SARS-CoV-2 in Clark County, there are not enough confirmed cases to make any scaling reasonable in regards to accurate inference and computational feasibility.

Ignoring computational limitations for a bit, the Lagrangian framework would theoretically work just as well for Clark County data as it did for the Diamond Princess. In fact, there are mainly two differences between the cruise ship data and the county data: (1) population size and (2) population transiency. Although the difference in population size poses an insurmountable problem, the population transiency can be easily remedied.

In the delineated model, N denoted the size of the studied population. Since there were very minor changes in the Diamond Princess' population, N was constant. However, Clark County certainly does not have a constant population. This county hosts a unique economy, heavily reliant on accommodation, food services, and retail trade. These industries are, in turn, dependent upon tourism, which is highly impressionable by a global pandemic. Typically, Las Vegas experiences a significant amount of tourism. In 2019, 42.5 million people visited the city [49]. So, the population in Clark County is augmented by tourists,

which are known to provide high risk in vector-borne diseases [50–52].

Let $\eta(t)$ be the number of tourists in Clark County at time t. To remain consistent in notation, it may be best to let t denote the number of days. Through this definition, $\eta(t)$ is a function of time and is likely sinusoidal in nature to match seasons or events. However, in the face of the SARS-CoV-2 pandemic, tourism dropped and remained minimal, revealing that $\eta(t)$ is likely to be quadratic or inverse-sigmoidal for the study period. If n is defined as Clark County's resident population, then $N(t) = n + \eta(t)$ denotes the current number of individuals in Clark County on day t.

With this altered definition of population size, there remains only one minor change to the Lagrangian framework: the inclusion and exclusion of certain agents each day. To keep the approach unsupervised, one would need to simulate each of the $N(t)$ agents each day, regardless of whether or not that agent is currently in Clark County. Suppose that the first n agents of the $N(t)$ population are the residents of Clark County and the remaining $\eta(t)$ agents are the tourists. For each of the tourists on each day, one can make a Bernoulli draw to determine if that agent is in Clark County or elsewhere. If that agent is not currently in Clark County, then their state of infection is not considered in the infection probability outlined in (1). This modification will enable the model to accurately analyze a transient population so long as a reasonable function $\eta(t)$ is selected. The simulated output for n_t should be comparable to the data gathered by reputable sources, such as the Southern Nevada Health District, so truly personalized inference would be obtained.

7 Conclusion

This thesis argues for the use of Lagrangian models when feasible to analyze epidemics so that one does not commit the ecological fallacy. To facilitate this argument, an ABM-surrogate model is provided that can track novel infectious diseases. This model is applied to the analysis of the SARS-CoV-2 outbreak aboard the Diamond Princess cruise ship, and

possible extensions were discussed when considering larger data sets, such as Clark County, Nevada.

In such a well-documented pandemic, it may seem that adequate data would render such a project trivial. However, this statistical model can provide personalized inference for sub-populations of a heterogeneous population by providing more accurate estimates for the population's parameters of interest, namely $\phi_{0,1,i,t}$. Additionally, the obtained inference can help construct guidelines for dealing with the next inevitable pandemic.

Further extensions of such a model could be the inclusion of individual characteristics, such as those used in the simulated study in Section 3. Then, researchers could obtain personalized inference dependent upon the potential inclusion of parameters like age, sex, working location, and more.

However, the most important work to be done in the field of Lagrangian inference is in scalability. Although the scalability problem is not unique to Lagrangian models, there lacks a regularization theory for ABMs. Much of the recent work on scalability in similar models requires sparsity or orthogonality in covariance matrices. These properties are rarely found in ABMs. A comprehensive treatment of scalability is not only desired for the sake of the Lagrangian perspective, but it is absolutely necessary to obtain personalized inference, which is required for the progress of personalized medicine, intervention, and treatment.

References

[1] P. J. Diggle and R. J. Gratton, "Monte Carlo methods of inference for implicit statistical models," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 46, no. 2, pp. 193–212, 1984.

[2] F. Hartig, J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth, "Statistical inference for stochastic simulation models–theory and application," *Ecology Letters*,

vol. 14, no. 8, pp. 816–827, 2011.

[3] E. Amouroux, S. Desvaux, and A. Drogoul, "Towards virtual epidemiology: an agent-based approach to the modeling of H5N1 propagation and persistence in North-Vietnam," in *Pacific Rim International Conference on Multi-Agents*, pp. 26–33, Springer, 2008.

[4] C. Siettos, C. Anastassopoulou, L. Russo, G. C, and M. E, "Modeling the 2014 Ebola virus epidemic-agent-based simulations, temporal analysis and future predictions for Liberia and Sierra Leone," *PLoS Currents*, vol. 1, no. 7, 2015.

[5] J. Parker, "A flexible, large-scale, distributed agent based epidemic model," in *2007 Winter Simulation Conference*, pp. 1543–1547, IEEE, 2007.

[6] B. D. Marshall and S. Galea, "Formalizing the role of agent-based modeling in causal inference and epidemiology," *American Journal of Epidemiology*, vol. 181, no. 2, pp. 92–99, 2015.

[7] L. Perez and S. Dragicevic, "An agent-based approach for modeling dynamics of contagious disease spread," *International Journal of Health Geographics*, vol. 8, no. 1, p. 50, 2009.

[8] A. M. El-Sayed, P. Scarborough, L. Seemann, and S. Galea, "Social network analysis and agent-based modeling in social epidemiology," *Epidemiologic Perspectives & Innovations*, vol. 9, no. 1, p. 1, 2012.

[9] A. H. Auchincloss and A. V. Diez Roux, "A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health," *American Journal of Epidemiology*, vol. 168, no. 1, pp. 1–8, 2008.

[10] J. B. Dunham, "An agent-based spatially explicit epidemiological model in MASON," *Journal of Artificial Societies and Social Simulation*, vol. 9, no. 1, 2005.

[11] K. M. Carley, D. B. Fridsma, E. Casman, A. Yahja, N. Altman, L.-C. Chen, B. Kaminsky, and D. Nave, "Biowar: scalable agent-based model of bioattacks," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 36, no. 2, pp. 252–265, 2006.

[12] J. M. Galán, A. López-Paredes, and R. Del Olmo, "An agent-based model for domestic water management in Valladolid metropolitan area," *Water Resources Research*, vol. 45, no. 5, 2009.

[13] N. E. Williams, M. L. O'Brien, and X. Yao, "Using survey data for agent-based modeling: design and challenges in a model of armed conflict and population change," in *Agent-Based Modelling in Population Studies*, pp. 159–184, Springer, 2017.

[14] A. Johnston, M. E. Hodson, P. Thorbek, T. Alvarez, and R. Sibly, "An energy budget agent-based model of earthworm populations and its application to study the effects of pesticides," *Ecological Modelling*, vol. 280, pp. 5–17, 2014.

[15] H. R. Parry, C. J. Topping, M. C. Kennedy, N. D. Boatman, and A. W. Murray, "A Bayesian sensitivity analysis applied to an agent-based model of bird population response to landscape change," *Environmental Modelling & Software*, vol. 45, pp. 104–115, 2013.

[16] V. Grimm and S. F. Railsback, "Agent-based models in Ecology: patterns and alternative theories of adaptive behaviour," in *Agent-Based Computational Modelling*, pp. 139–152, Springer, 2006.

[17] P. Patlolla, V. Gunupudi, A. R. Mikler, and R. T. Jacob, "Agent-based simulation tools in computational epidemiology," in *International Workshop on Innovative Internet Community Systems*, pp. 212–223, Springer, 2004.

[18] E. Hunter, B. Mac Namee, and J. D. Kelleher, "A taxonomy for agent-based models in human infectious disease epidemiology," *Journal of Artificial Societies and Social Simulation*, vol. 20, no. 3, 2017.

[19] P. M. Torrens, "Agent-based models and the spatial sciences," *Geography Compass*, vol. 4, no. 5, pp. 428–448, 2010.

[20] S. Arifin, R. R. Arifin, D. D. A. Pitts, M. S. Rahman, S. Nowreen, G. R. Madey, and F. H. Collins, "Landscape epidemiology modeling using an agent-based model and a geographic information system," *Land*, vol. 4, no. 2, pp. 378–412, 2015.

[21] Y. Yang, A. V. D. Roux, A. H. Auchincloss, D. A. Rodriguez, and D. G. Brown, "A spatial agent-based model for the simulation of adults' daily walking within a city," *American Journal of Preventive Medicine*, vol. 40, no. 3, pp. 353–361, 2011.

[22] B. Herd, *Statistical runtime verification of agent-based simulations*. PhD thesis, King's College London, 2015.

[23] K. S. Perumalla and B. G. Aaby, "Data parallel execution challenges and runtime performance of agent simulations on GPUs," *SpringSim*, vol. 8, pp. 116–123, 2008.

[24] B. G. Aaby, K. S. Perumalla, and S. K. Seal, "Efficient simulation of agent-based models on multi-GPU and multi-core clusters," in *Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques*, pp. 1–10, 2010.

[25] H. Baumgaertel and U. John, "Combining agent-based supply net simulation and constraint technology for highly efficient simulation of supply networks using APS systems," in *Winter Simulation Conference*, vol. 2, pp. 1765–1773, 2003.

[26] D. Pawlaszczyk, "Scalable multi agent based simulation-considering efficient simulation of transport logistic networks," in *12th ASIM Conference-Simulation in Production and Logistics*, Citeseer, 2006.

[27] M. Hybinette, E. Kraemer, Y. Xiong, G. Matthews, and J. Ahmed, "Sassy: a design for a scalable agent-based simulation system using a distributed discrete event infrastructure," in *Proceedings of the 2006 Winter Simulation Conference*, pp. 926–933, IEEE, 2006.

[28] M. Hooten, C. Wikle, and M. Schwob, "Statistical implementations of agent-based demographic models," *International Statistical Review*, vol. 88, no. 2, pp. 441–461, 2020.

[29] C. J. Topping, T. T. Høye, and C. R. Olesen, "Opening the black box—development, testing and documentation of a mechanistically rich agent-based model," *Ecological Modelling*, vol. 221, no. 2, pp. 245–255, 2010.

[30] "The most typical person in the world," *SBS Popular Science*, pp. 1–3, 2013. SBS News.

[31] "Cruise ship accounts for more than half of virus cases outside China - as it happened," *The Guardian Breaking News*, pp. 1–2, 2020. The Guardian.

[32] "About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port," *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[33] "About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (pt. 2)," *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[34] "About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (4th report)," *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[35] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (5th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[36] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (6th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[37] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (8th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[38] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (9th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[39] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (10th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[40] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (11th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[41] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (12th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[42] “About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (13th report),” *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[43] "About the new coronavirus infection confirmed on the cruise ship that called at Yokohama Port (14th report)," *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[44] "About the new coronavirus infection confirmed on the cruise ship being quarantined at Yokohama Port (announced on February 26)," *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[45] "About the PCR test results for new coronavirus infections related to passengers and crew of cruise ships being quarantined at Yokohama Port," *National Covid-19 Reports*, pp. 1–2, 2020. Ministry of Health, Labour, and Welfare - Japan.

[46] "If you've been exposed to the coronavirus," *Harvard Health Publishing*, pp. 1–2, 2020. Harvard Medical School.

[47] "Mortality risk of Covid-19," *Our World in Data - Weekly*, pp. 1–5, 2020. Our World in Data.

[48] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[49] "Slight gains in Vegas tourism numbers for 2019," *Weekly Reports*, pp. 1–3, 2020. Travel Weekly.

[50] S. Ninphanomchai, C. Chansang, Y. L. Hii, J. Rocklöv, and P. Kittayapong, "Predictiveness of disease risk in a global outreach tourist setting in Thailand using meteorological data and vector-borne disease incidences," *International Journal of Environmental Research and Public Health*, vol. 11, no. 10, pp. 10694–10709, 2014.

[51] E. B. Hayes, "Looking the other way: preventing vector-borne disease among travelers to the United States," *Travel Medicine and Infectious Disease*, vol. 8, no. 5, pp. 277–284, 2010.

[52] T. D. Vermeulen, J. Reimerink, C. Reusken, S. Giron, and P. J. de Vries, "Autochthonous Dengue in two Dutch tourists visiting département var, Southern France, July 2020," *Eurosurveillance*, vol. 25, no. 39, pp. 176–201, 2020.