Bias and Bias-Correction for Individual-Level Models of Infectious Disease

February 14, 2020

Abstract

Accurate infectious disease models can help scientists understand how an ongoing disease epidemic spreads and forecast the course of epidemics more effectively. Considering various factors that affect the spread of a disease (e.g. geographical, social, domestic, and genetic), a class of individual-level models (ILMs) was developed to incorporate population heterogeneity. In these models, inferences are developed within a Bayesian Markov chain Monte Carlo (MCMC) framework, obtaining posterior estimates of model parameters. The issues of bias of parameter estimates, and methods for bias correction, have been widely studied with respect to many of the most established and commonly used statistical models and associated methods of parameter estimation. However, these methods are not directly applicable to infectious disease data. This paper investigates circumstances in which ILM parameter estimates may be biased in some simple disease system scenarios. Further, we aim to compare the performance of bias-corrected estimates of ILM parameters, using simulation, with the posterior estimates of the parameter. We also discuss the factors that affect the performance of these estimators.

Keywords: infectious disease model, disease epidemic, individual-level models (ILMs), Markov chain Monte Carlo (MCMC), parameter estimation, posterior estimate, bias, bias correction

1 Introduction

Infectious diseases have always been part of human history. Through the ages, humankind has struggled to survive and control these diseases. Although ecological and public health sciences have tremendously improved, we are still struggling with common contagious diseases such as influenza or HIV (human immunodeficiency virus). Many questions related to the transmission dynamics of infectious disease remained unanswered. Such questions include how rapidly a disease spreads through a population over time, and what the best policy is for taking control over the disease. These are major concerns of national and international policy makers (Riley et al., 2014).

Mathematical models have been widely used to try to answer these questions and help policy makers. Mathematical modeling and simulation allow for rapid assessment of the spread of a disease. However, a

useful model depends on the assumptions embedded therein and the degree the model is applicable . For instance, a heterogeneous population assumption can greatly improve model accuracy compared with a homogeneous mixing because it takes into account more real-world factors in the model (Savill et al., 2006). We can use these models retrospectively to indicate how diseases were transmitted (e.g., Malik et al., 2014; Cauchemez et al., 2011). We can also use them prospectively to inform policy makers planning different interventions to control the spread of a disease (e.g., Tildesley et al., 2006). When modeling, we will also likely be interested in successfully identifying risk factors associated with transmission.

In general, infectious disease models have been developed inside various frameworks, e.g., (homogeneous mixing) compartmental, metapopulation, individual-level, spatial, and network frameworks. The choice of framework depends on the nature of the disease and population, the availability of data, and the purpose of the modeling (e.g., O'Neill, 2010; Bifolchi et al., 2013; Deardon et al., 2010).

The most common framework is the homogeneous compartmental framework. In this framework, at any point in time individuals can be classified into a series of groups based upon various stages in the life cycle of the disease. A set of equations can then be developed that models the rates of transitions between the groups over time. Typical options for categorizing individuals are as follows:

- Susceptible (S) category: individuals capable of contracting the disease but not yet exposed;
- Exposed (E) category: individuals exposed to the disease but not yet capable of spreading it;
- Infectious (I) category: individuals infected and capable of spreading it;
- Removed (R) category: individuals unable to spread the disease. This can occur for a number of reasons, such as recovery with acquired immunity, quarantine, or death.

Each individual can only be in one category at any given point in time. Based on these categories, we can have different compartmental models: SEIRS, SIR, SI, and SIS.

In a metapopulation framework, epidemic transmission dynamics are affected by population structure. For instance, a population can be divided up between males and females, different age groups, or different locations. Models in this framework keep track of the number of individuals in each group. Dynamics of local populations are considered in an abbreviated way. This concept assumes a homogeneous mixing in discrete patches with separate parameters detailing transmission (or migration), between patches (Siettos and Russo, 2013).

However, in real life we generally do not have homogeneous and homogeneously mixing populations either in the population overall or in patches within the population. For example, variations in human populations exist because of geographical, social, domestic, and employment networks, genetics, different levels of immunity, and individual lifestyle factors. To model such factors, we need to turn our attention to individual-based or individual-level models. Deardon et al. (2010) introduced a class of discrete-time individual-level models (ILMs) that incorporate population heterogeneity. These models take into consideration various individual-level risk factors to model the transmission rate of disease for individuals over time. For example, disease

transmission through a population can be modeled through a binary or weighted contact network, spatial distance infection kernel, or combination of both.

To develop inference for such infectious disease models, we generally use a Bayesian Markov chain Monte Carlo (MCMC) framework (e.g., Gamerman and Lopes, 2006), obtaining posterior estimates of model parameters. The ability to handle missing data, as well as complex models, makes the Bayesian MCMC framework highly suitable for infectious disease model inference. However, MCMC methods can be computationally intensive, especially for large population sizes, complex models, and/or large amounts of missing/latent information. Different algorithms can be used to construct a Markov chain with a stationary distribution identical to the posterior distribution. The Metropolis-Hastings (MH) algorithm is the most commonly used.

In any statistical problem, parameters we estimate can have a tendency to exhibit bias depending on the estimation procedure and nature of the system being studied. If we can estimate these biases, then we can use them to bias-correct an estimator. The non-parametric bootstrap algorithm can be easily adapted to give estimates of bias in cases where we have independent data. Jackknife estimates of bias can also be useful. However, such corrections can be dangerous due to larger variance in the bias-corrected estimator and so should be done with caution (Efron and Tibshirani, 1993).

Of course, in infectious disease systems it is generally not sensible to bootstrap the data as dependence relationships therein will be lost in the bootstrap samples. Further, defining residuals to bootstrap, for say a spatio-temporal system, also does not appear to be an obvious or sensible way to proceed. However, we can consider the use of parametric bootstrap under which our "bootstrapped" data are simulated by the model.

In this paper, we aim to quantify the bias that may occur in disease systems when fitting disease transmission models to data. We also examine if a parametric bias correction method can lead to improved estimation. The paper is laid out as follows. Section 2 gives brief background information about ILMs and the type of ILM that is to be used for the rest of this paper. We also discuss bias in ILMs for discrete-time ILMs. In Section 3, we introduce our bias-correction method and investigate its performance through simulation studies, and show the results. In Section 4, we go on to summarize the results of the simulation studies and discuss possible future directions for research.

2 ILMs of Disease Transmission

ILMs of disease transmission are designed to model how an infectious disease spreads through a heterogeneous population. These models define disease infection or exposure events and are placed within a so-called compartmental (multi-state) framework that defines the progress of individual-level status over time. For the purpose of simplicity, we consider the simple two-state susceptible-infectious (SI) framework for ILMs. In this framework, individuals begin in the susceptible state (S) and if/when infected become immediately infectious (I) and remain in this state indefinitely. These models take into account individual-level risks such as spatial distance between individuals. $S(t)$ and $I(t)$ represent the sets of susceptible and infectious

individuals at time t, respectively. Individuals must be in one of these two sets at any given time, and cannot be in more than one at any given time. These two sets define the epidemic event history for $t = 1, \ldots,$ $t = t_{max}$, where t=1 is the time when the first infection is observed and $t = t_{max}$ is when the epidemic ends. We define the infection rate λ_t as a rate of transition of individuals from $S \to I$ at time t. In classical SI models, where the population mixes homogeneously and we have a homogeneous population, the infection rate is defined as

$$\lambda_t = \lambda |S(t)||I(t)|, \tag{1}$$

where λ is an infectivity rate parameter. However, such models do not account for heterogeneities in the population. In ILM-SI models, the infection rate of susceptible individual i at time t is given by

$$\lambda_{it} = \Omega_S(i)\Big\{ \sum_{j \in I(t)} \Omega_T(j)\kappa(i,j)\Big\} + \epsilon(i,t), \tag{2}$$

where $\Omega_S(i)$ is a susceptibility function representing risk factors associated with a given susceptible individual i contracting the disease; $I(t)$ is the set of individuals infectious at time t; $\Omega_T(j)$ is a transmissiblity function representing risk factors associated with the transmission of the disease from a given infectious individual j; $\kappa(i,j)$ is an infection kernel that represents risk factors jointly associated with specific pairs of susceptible and infectious individuals; and $\epsilon(i,t)$ is a random sparks function representing infections that are not well explained by $\Omega_S(i)$, $\Omega_T(j)$, or $\kappa(i,j)$ (Deardon et al., 2010).

Typically, $\Omega_S(i)$ and $\Omega_T(j)$ are linear functions of covariates representing genetic or environmental risk factors, or possibly vaccination status. $\Omega_S(i)$, which accommodates potential risk factors associated with susceptible individual i contacting the disease, is treated as a linear function of the covariates, i.e., $\Omega_S(i) = a_0 + a_1X_1(i) + a_2X_2(i) + \cdots + a_{n_s}X_{n_s}(i)$, where $X_1(i), X_2(i), \cdots, X_{n_s}(i)$ denote the n_s covariates associated with susceptible individual i and the susceptibility parameters $a_0, \cdots, a_{n_s} > 0$. The sparks term, ϵ, represents infections originating from outside the population being observed or some other poorly modeled infection mechanism (Kwong and Deardon, 2012).

The general discrete-time ILM of Deardon et al. (2010), which describes the probability of a susceptible individual, i, entering the exposed state at discrete time t+1, is defined as $P_{it} = 1 - \exp(-\lambda_{it})$. The infection kernel, $\kappa(i,j)$, could be a function of distance, or possibly an indicator function denoting the presence or absence of an edge in a graph representing some type of contact network through which the disease can potentially spread. The random sparks term typically represents infection coming in from outside the study population.

The likelihood function for the discrete-time ILM is given by

$$L(S, I|\theta) = \prod_{t=0}^{t_{max}} \left[\prod_{i \in I(t+1) \setminus I(t)} \Big\{1 - \exp(-\lambda_{it})\Big\} \right] \left[\prod_{i \in S(t+1)} \exp(-\lambda_{it}) \right], \tag{3}$$

where θ is a vector of unknown parameters, t_{max} is the maximum infectious time, $I(t+1) \setminus I(t)$ is the set of infected individuals appearing newly infectious at time $t+1$, and $S(t+1)$ is the set of individuals still susceptible at time $t+1$.

In this manuscript, we consider two specific ILMs: a spatial model and a homogeneous (or saturated network) model. In the spatial ILM considered here, $\kappa(i,j)$ is an Euclidean distance-based power-law spatial kernel. No covariates or spark infectious term are included in this model, so we set $\Omega_S(i) = \alpha$, $\Omega_T(j) = 1$, and $\epsilon(i,t) = \epsilon = 0$. Thus, the probability of a susceptible individual, i, entering the exposed state at time $t+1$ is defined as $P_{it} = 1 - \exp\Big[-\alpha \sum_{j \in I(t)} d_{ij}^{-\beta} \Big]$, where d_{ij} denotes the Euclidean distance between individuals i and j. The homogeneous model we consider here has $\kappa(i,j) = 1$, $\forall i,j$, along with $\Omega_S(i) = \alpha$, $\Omega_T(j) = 1$ and $\epsilon = 0$. Thus, the probability becomes $P_{it} = 1 - \exp\Big[-\alpha \sum_{j \in I(t)} 1 \Big]$.

Now, we can use a Bayesian framework to estimate model parameters θ using the posterior distribution that is given by Bayes' theorem, as follows:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\ \pi(\theta)}{\pi(D)}, \tag{4}$$

where $\pi(D|\theta)$ is the likelihood function, $\pi(\theta)$ is the prior distribution of θ, and $\pi(D) = \int \ell(D|\theta)\ p(\theta)\ d\theta$ is a normalization constant. Because $\pi(\theta)$ is not available in closed form, we resort to MCMC methods to generate samples from the posterior and carry out Monte Carlo inference. Here, we use the EpiILM R package to simulate data and fit discrete-time ILMs within a Bayesian framework using Metropolis-Hastings MCMC (Warriyar and Deardon, 2018). The R package EpiILMCT is used to do the same for the continuous-time ILMs(Almutiry et al., 2018).

2.1 Bias in ILMs

By definition, bias of an estimator $\widehat{\theta}$ of parameter θ is defined as

$$B[\ \widehat{\theta}\ ] = \mathbb{E}[\ \widehat{\theta}\ ] - \theta, \tag{5}$$

In ILMs, parameter estimators are the posterior mean estimates from a Bayesian MCMC analysis with arbitrary chosen vague priors since there is no prior information about the parameters distribution. We therefore define

$$\widehat{B}[\ \widehat{\theta}\ ] = \mathbb{E}[\ \widehat{\theta}_\pi\ ] - \theta \tag{6}$$

where $\widehat{\theta}_\pi$ is the posterior mean for a single dataset and the expectation is taken over repeated datasets. We refer to $\mathbb{E}[\widehat{\theta}_\pi]$ as the posterior bias. In this paper, we will investigate the posterior bias for two types of ILMs: homogeneous and spatial. For the homogeneous case, we investigate both discrete- and continuous-time frameworks. For the spatial case, we investigate the discrete-time framework.

3 Simulation Scenarios for Bias in SI Models

To investigate the posterior bias for ILMs, we consider three scenarios based on the two types of ILMs. We want to investigate bias in discrete- and continuous-time homogeneous and discrete-time spatial models for model parameters explained by equation 6. First, we simulate epidemic data for the scenarios. An epidemic simulation starts with randomly chosen initial infective individual k at time $t_0 = 0$ where $k = 1, \cdots, n$ and n is the population size. Then, waiting times, Ws, are calculated for each susceptible individual and an individual with minimum waiting time is chosen as the next infected individual, with $I_{t+1} = I_t + W$. This process is repeated until no infective remains in the population.

3.1 Bias Simulation in SI Homogeneous Models

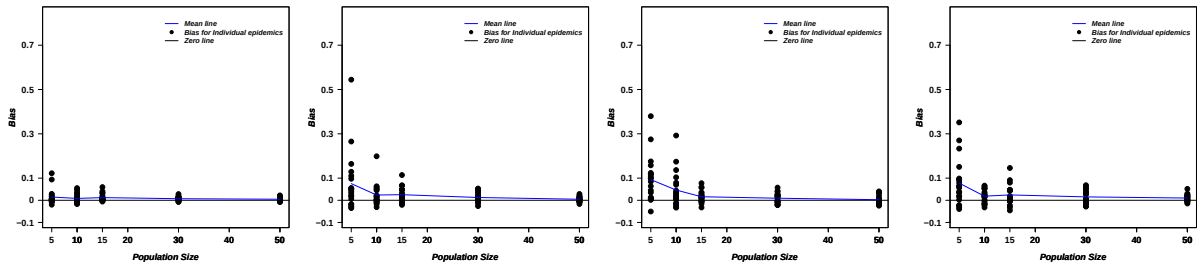
Homogeneous mixing models assume that all susceptible individuals have the same rate of getting infected at any given point in time. Thus, a homogeneous model contains only one parameter: the susceptibility parameter, α. To examine the bias for this parameter in these models, we fitted the models within a Bayesian framework via the MCMC method for cases of both discrete- and continuous-time, for different population sizes and values of the parameter α. We designed our simulation study based on 20 epidemics. The MCMC chains were run for 50000 MCMC iterations per epidemic. We discard 10% of the MCMC iterations for each epidemic as burn-in. Values of α between 0.05 and 0.2 with a step size of 0.025 for five population sizes (5, 10, 15, 30, and 50 individuals) were considered. We also chose a prior with a vague uniform (0, 10000) distribution for α.

3.1.1 Bias in SI Homogeneous Discrete-Time Models

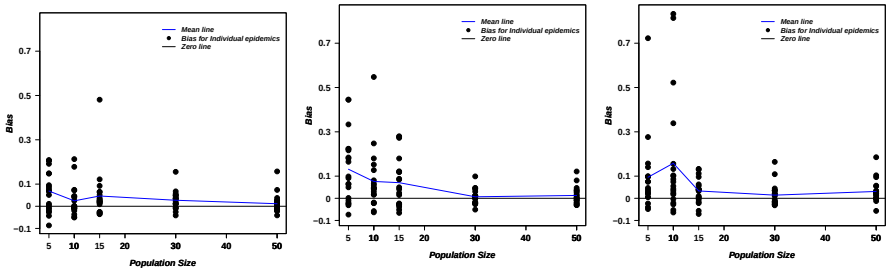
Figure 1 shows the posterior mean bias, $B[\alpha]$, in a homogeneous discrete-time framework for different values of model parameter α for different population sizes. As we can see, posterior mean bias is positive, for the population sizes and parameter values considered, which means we tend to overestimate α. We can also see that both population size and the value of α affect the bias, with the overestimation of the parameter tending to be higher for small population sizes (n=5 and n=10) and for larger values of α. In addition, the variance of the posterior estimates between epidemics seems to be much higher in smaller populations than for larger ones, as we would expect.

3.1.2 Bias in Homogeneous Continuous-Time SI Models

Figure 2 shows the posterior mean bias, $B[\alpha]$, in a homogeneous continuous-time framework for different values of model parameter α for different population sizes. The results are consistent with those above for the discrete-time framework, in terms of having a positive posterior mean bias and the same effects of

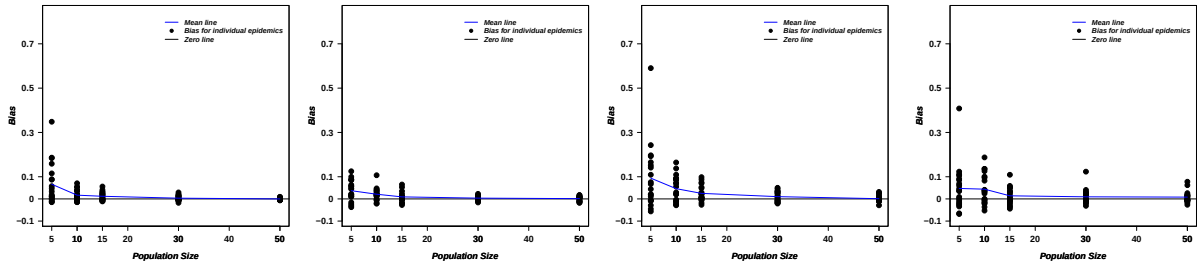


(a) $\alpha = 0.05$ (b) $\alpha = 0.075$ (c) $\alpha = 0.1$ (d) $\alpha = 0.125$

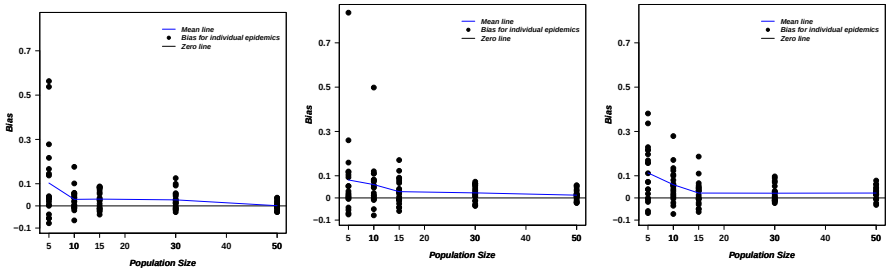


(e) $\alpha = 0.15$ (f) $\alpha = 0.175$ (g) $\alpha = 0.2$

Figure 1: Posterior bias of parameter α versus population size in a discrete-time homogeneous model. Each plot shows the result for a different α value.



(a) $\alpha = 0.05$ (b) $\alpha = 0.075$ (c) $\alpha = 0.1$ (d) $\alpha = 0.125$



(e) $\alpha = 0.15$ (f) $\alpha = 0.175$ (g) $\alpha = 0.2$

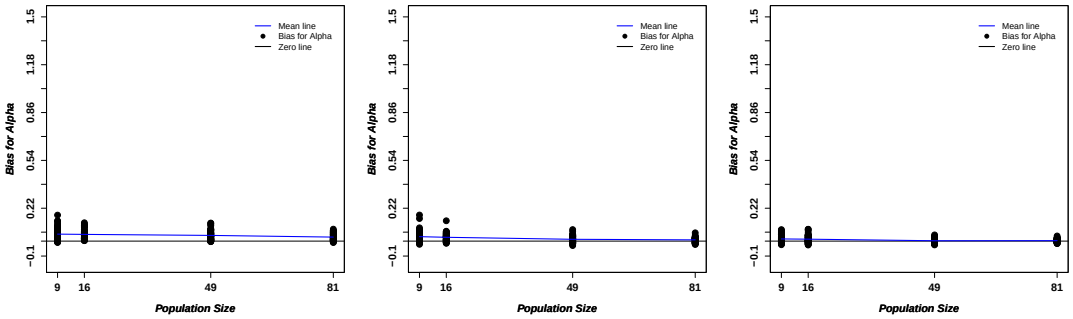
Figure 2: Posterior bias of parameter α versus population size in a continuous-time homogeneous model. Each plot shows the result for a different α value.

population size and value of α on bias. Once again, bias for small population sizes tends to be higher than for larger populations. In addition, as the value of α increases we observe more bias in the posterior mean estimates of α. However, variance of posterior mean estimates over different datasets seems to be lower than in estimations for discrete-time models.

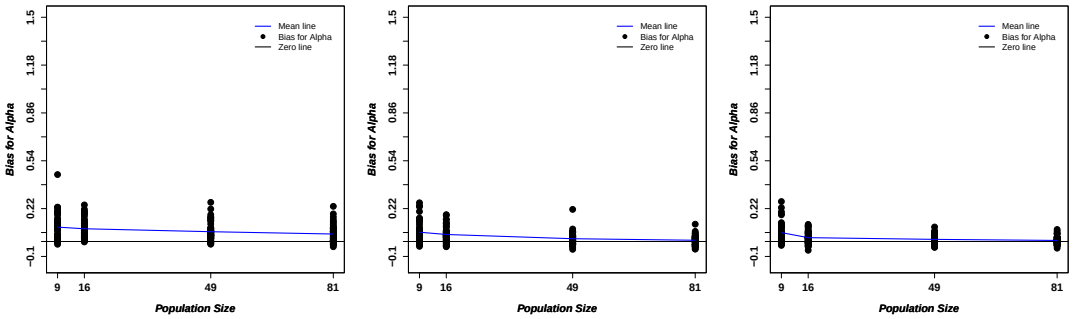
3.2 Bias in Spatial SI Models

Spatial models assume the probability of disease transmission at any given time point is based on distance. This means the probability of an individual becoming infective is calculated based on the distance between the individual and other infective individuals. In this case, our model includes an extra spatial parameter, β, in addition to α, the susceptibility parameter. Here, we assume individuals are evenly distributed in space across a regular grid one unit apart along the x and $y$axes. Thus, we have the populations laid out on an $n \times n$ lattice.

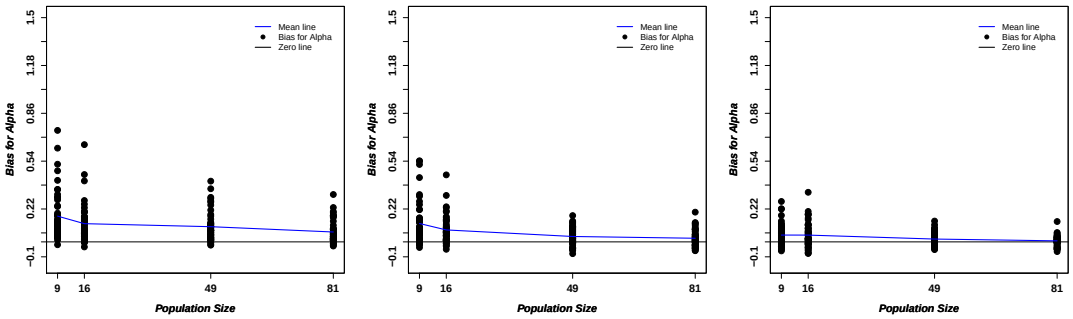
To estimate parameters in these models, we design a simulation study based on 50 epidemics and 100000 MCMC iterations per epidemic. Values of α are between 0.05 and 0.2 with a step size of 0.025 and values of β are between 0.5 and 2.5 with a step size of 0.5, for four population sizes of 9 ($n = 3$), 16 ($n = 4$), 49 ($n = 7$), and 81 ($n = 9$) individuals with a vague prior distribution.



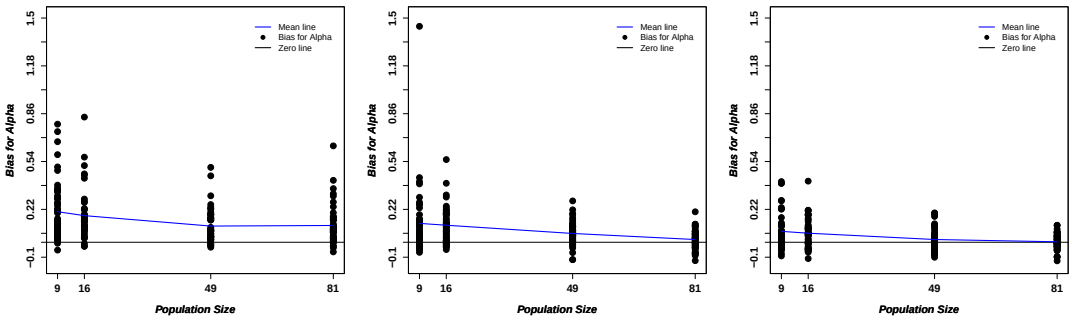
(a) $\alpha = 0.05$, $\beta = 0.5$ (b) $\alpha = 0.05$, $\beta = 1.5$ (c) $\alpha = 0.05$, $\beta = 2.5$



(d) $\alpha = 0.1$, $\beta = 0.5$ (e) $\alpha = 0.1$, $\beta = 1.5$ (f) $\alpha = 0.1$, $\beta = 2.5$

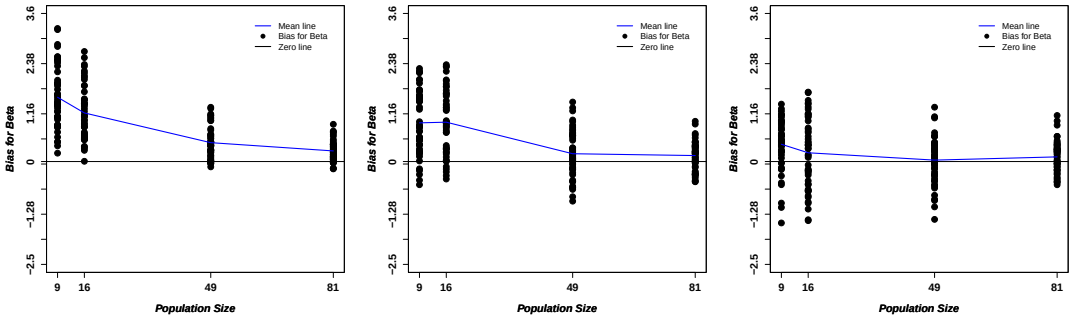


(g) $\alpha = 0.15$, $\beta = 0.5$ (h) $\alpha = 0.15$, $\beta = 1.5$ (i) $\alpha = 0.15$, $\beta = 2.5$

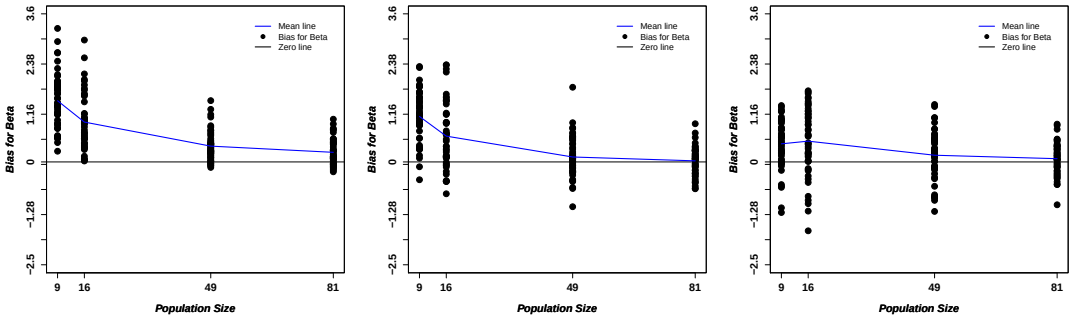


(j) $\alpha = 0.2$, $\beta = 0.5$ (k) $\alpha = 0.2$, $\beta = 1.5$ (l) $\alpha = 0.2$, $\beta = 2.5$

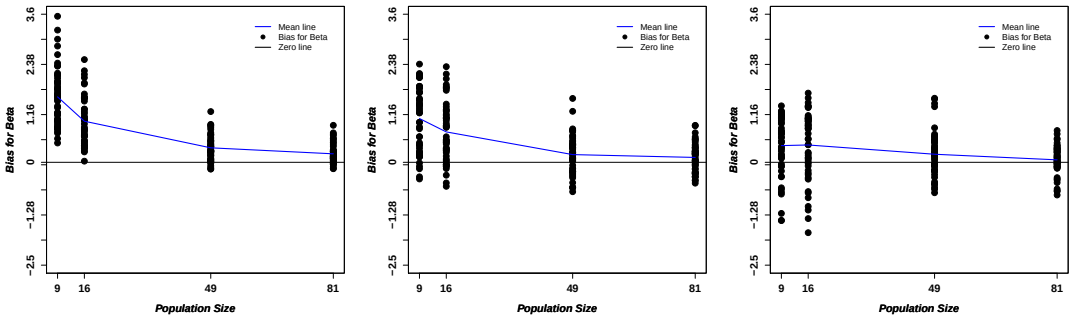
Figure 3: Posterior bias of α versus population size for a discrete-time spatial model for varying α and β.



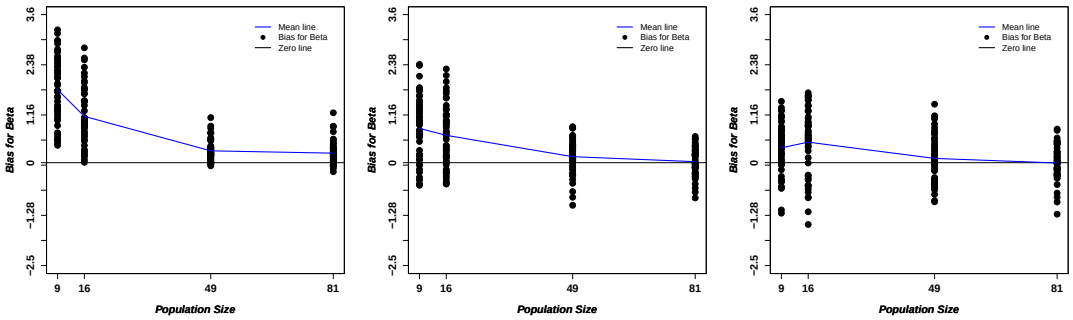
(a) $\alpha = 0.05$, $\beta = 0.5$ (b) $\alpha = 0.05$, $\beta = 1.5$ (c) $\alpha = 0.05$, $\beta = 2.5$



(d) $\alpha = 0.1$, $\beta = 0.5$ (e) $\alpha = 0.1$, $\beta = 1.5$ (f) $\alpha = 0.1$, $\beta = 2.5$



(g) $\alpha = 0.15$, $\beta = 0.5$ (h) $\alpha = 0.15$, $\beta = 1.5$ (i) $\alpha = 0.15$, $\beta = 2.5$



(j) $\alpha = 0.2$, $\beta = 0.5$ (k) $\alpha = 0.2$, $\beta = 1.5$ (l) $\alpha = 0.2$, $\beta = 2.5$

Figure 4: Posterior bias of β versus population size for a discrete-time spatial model for varying α and β.

3.2.1 Bias in Spatial Discrete-Time SI Models

Figure 3 shows the posterior mean bias $B[\alpha]$, and Figure 4 shows the posterior mean bias $B[\beta]$, for varying values of α and β for population sizes of 9, 16, 49, and 81 individuals modeled in a discrete-time SI compartmental framework. MCMC for both parameters shows good convergence. These figures show that posterior mean bias tends to be positive, in this scenario, for both parameters, α and β. This means we have more overestimation than underestimation of parameters. The mean bias estimates for both parameters tend to be higher for smaller population sizes than for larger population sizes. We also see that bias tends to be larger and that more variability exists between epidemic datasets for smaller values of both β and, especially, α.

4 Bias Correction for ILMs

We now turn to the issue of whether we can correct for the bias identified in the ILM parameters. We do this using a regression-based approach focusing on the two-parameter discrete-time spatial models. The general idea is to simulate multiple epidemics across a grid of (α, β) pairs for a given population. We then calculate the mean bias at each of the (α, β) points in the grid. We then regress the mean bias upon predictors $\mathbb{E}[\ \widehat{\alpha_\pi}\ ]$, $\mathbb{E}[\ \widehat{\beta_\pi}\ ]$, $\mathbb{E}^2[\ \widehat{\alpha_\pi}\ ]$, and $\mathbb{E}^2[\ \widehat{\beta_\pi}\ ]$. Specifically, the bias of posterior mean estimates (PME) of α is modeled as

$$\widehat{B}[\ \widehat{\alpha}\ ] = a_0 + a_1\,\mathbb{E}[\ \widehat{\alpha_\pi}\ ] + a_2\,\mathbb{E}^2[\ \widehat{\alpha_\pi}\ ] + a_3\,\mathbb{E}[\ \widehat{\beta_\pi}\ ] + a_4\,\mathbb{E}^2[\ \widehat{\beta_\pi}\ ], \tag{7}$$

and the bias of PME of β is modeled as

$$\widehat{B}[\ \widehat{\beta}\ ] = b_0 + b_1\,\mathbb{E}[\ \widehat{\alpha_\pi}\ ] + b_2\,\mathbb{E}^2[\ \widehat{\alpha_\pi}\ ] + b_3\,\mathbb{E}[\ \widehat{\beta_\pi}\ ] + b_4\,\mathbb{E}^2[\ \widehat{\beta_\pi}\ ], \tag{8}$$

where $\mathbb{E}_\pi[\cdot]$ denotes the posterior mean estimate and $\widehat{B}[\cdot]$ denotes the mean bias from the parameter. We can attempt to use equations 7 and 8 to correct for the bias and investigate the performance of the quadratic regression bias correction method in terms of improving the point estimates of the parameters of the spatial ILM.

We then define the bias-corrected estimate (BCE) of a parameter to be the difference between the posterior mean estimate and estimated value of $\widehat{B}$ through the quadratic regression model:

$$BCE[\ \cdot\ ] = \mathbb{E}[\ \widehat{\cdot}_\pi\ ] - \widehat{B}[\ \widehat{\cdot}\ ], \tag{9}$$

We also then define the bias-corrected estimate bias (BCB) as the difference between the bias-corrected

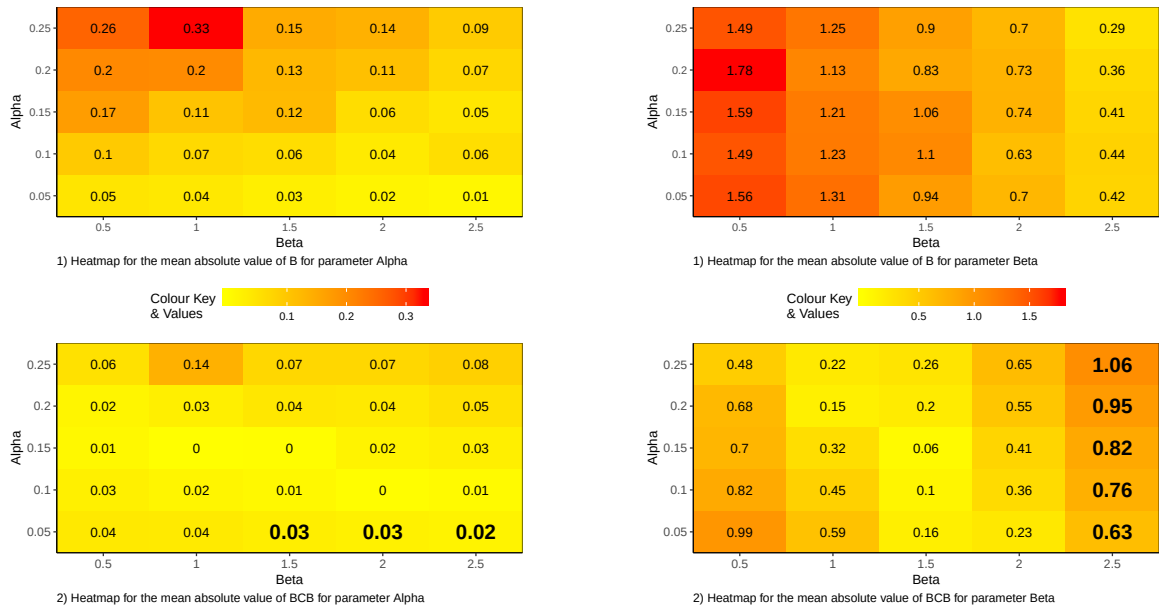
estimate of a parameter and the true parameter:

$$BCB[\,\cdot\,] = BCE[\,\cdot\,] - [\,\cdot\,], \tag{10}$$

We also use the notations $|B[\,\cdot\,]|$ and $|BCB[\,\cdot\,]|$ to refer to the absolute bias and absolute corrected bias of a parameter, respectively. These are used in ascertaining performance of the bias-correcting method as the magnitude of the bias is more important than the direction (i.e., over/underestimation) for our purposes. To ensure that using BCE is a safe way to improve the bias for parameter estimators, $\mathbb{E}[\,\widehat{\cdot}_\pi\,]$, we checked the variability in the BCE. For almost all cases of α and β, BCE shows a smaller variability than $\mathbb{E}[\,\widehat{\cdot}_\pi\,]$. Thus, we can safely use this bias-correction method for our purposes.

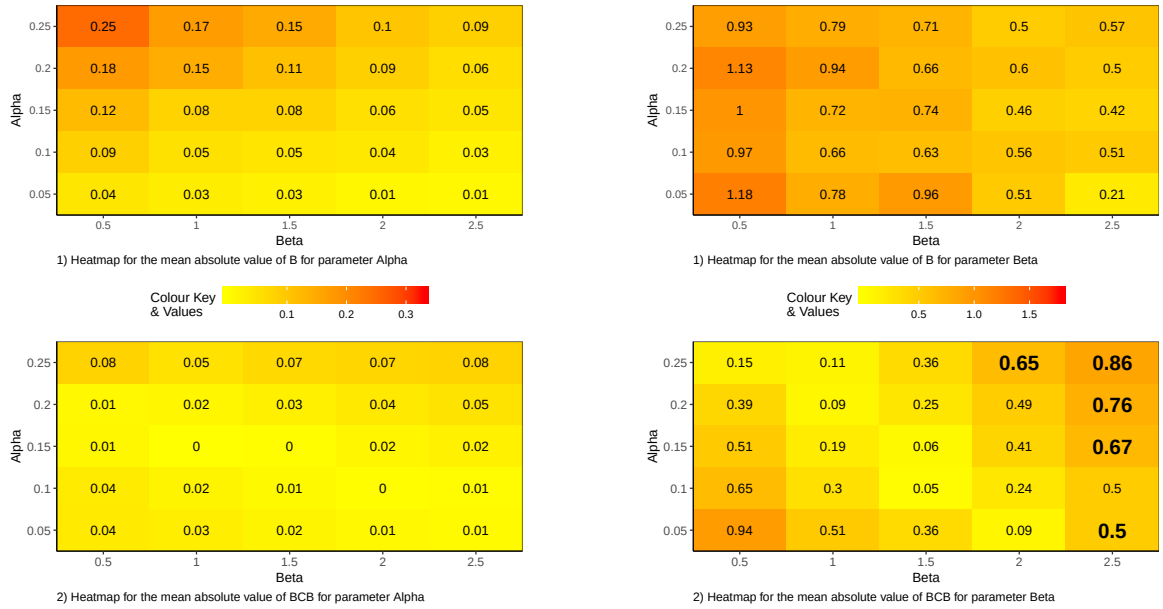
4.1 Simulation Scenarios for BCB in ILMs

The regression bias correction method is considered here for population sizes of 9, 16, 49, and 81 individuals, with these individuals placed on a regular grid 1 unit apart along the x and $y$axes as described in Section 3.2. The model considered here is the spatial ILM within an SI framework. We consider an isolated population of individuals distributed on an $n \times n$ lattice (n = 3, 4, 7, and 9). We calculate $|B[\cdot]|$ and $|BCB[\cdot]|$ for both parameters α and β, averaged over 50 epidemic replicates, over a grid defined by pairs of $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ and $\beta \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$.



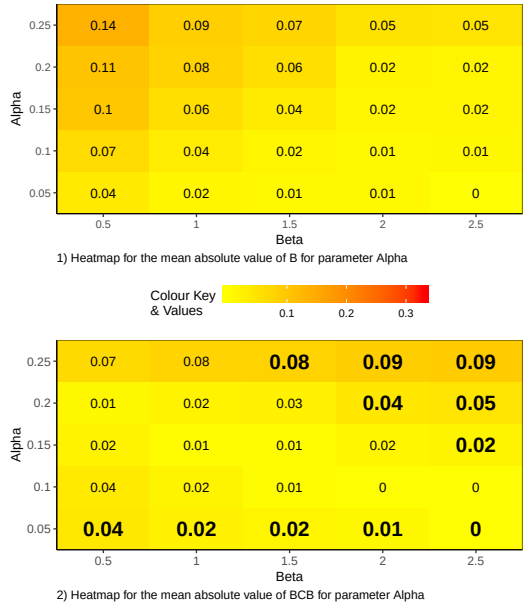
(a) Heat maps for mean absolute values of B and BCB for α (b) Heat maps for mean absolute values of B and BCB for β

Figure 5: Comparison between mean absolute value of B and BCB for parameter α and β based on different values of parameters (population size n = 9)

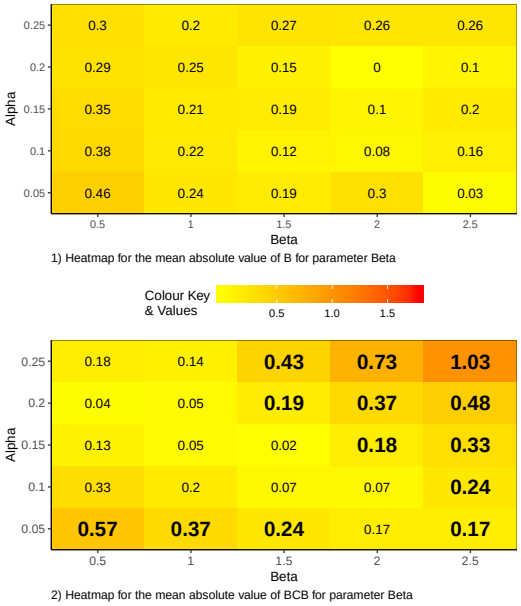


(a) Heat maps for absolute values of B and BCB for α (b) Heat maps for mean absolute values of B and BCB for β

Figure 6: Comparison between mean absolute value of B and BCB based on different values of parameters α and β (population size n = 16)

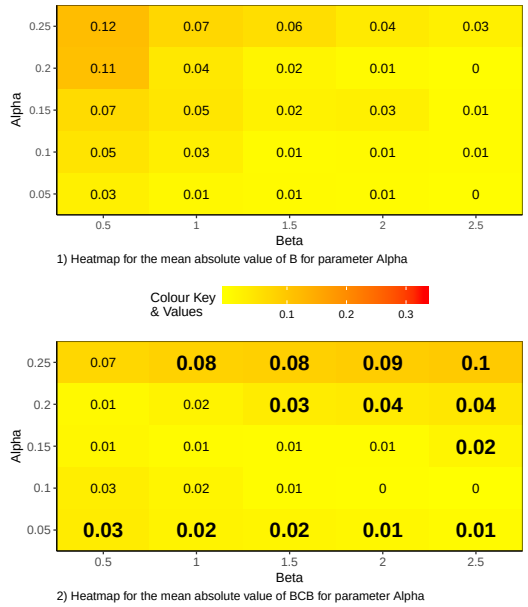


(a) Heat maps for absolute values of B and BCB for α

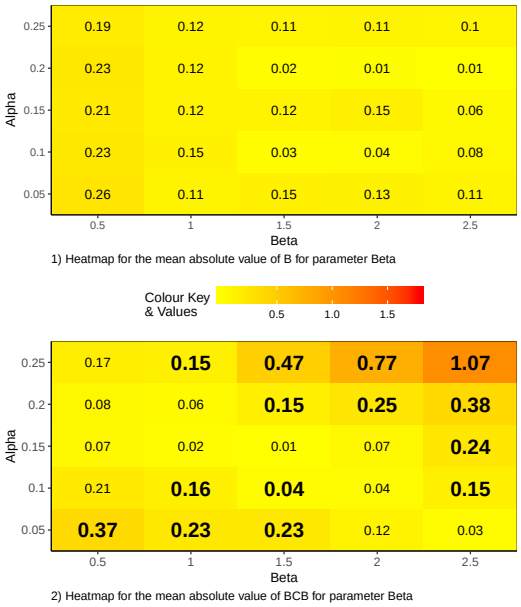


(b) Heat maps for mean absolute values of B and BCB for parameter β

Figure 7: Comparison between mean absolute value of B and BCB based on different values of parameters α and β (population size n = 49)



(a) Heat maps for absolute values of B and BCB for α



(b) Heat maps for mean absolute values of B and BCB for parameter β

Figure 8: Comparison between mean absolute value of B and BCB based on different values of parameters α and β (population size n = 81)

4.2 Results from Simulation Scenarios

Figures 5a, 6a, 7a, and 8a and Figures 5b, 6b, 7b, and 8b show the absolute mean posterior bias $|B[\cdot]|$ and mean bias corrected $|BCB[\cdot]|$ estimates of the model parameters, α and β, respectively, for the four population sizes.

As seen in Section 3, posterior bias for α and β tends to increase as the population size decreases. For both parameters, the highest value for $|B[\cdot]|$ occurs when the population size is 9. For small populations (e.g., n = 9 or 16), in which $|B[\cdot]|$ tends to be large, the bias correction method reduces bias and leads to improved posterior point estimates for both parameters α and β. For large populations, in which $|B[\cdot]|$ tends to already be small, the bias-correction method can increase bias, especially around the edges of the simulation parameter space. This implies that the bias-correction method.This would implies that the bias correction method probably should not be used in situations with small amounts of bias.

For a population size of 9, we observe that the absolute bias of α is smaller under the bias-correction method for all values of α and β, except when the population size is increasing, α has its smallest value (0.05) and β is larger than 1.5 (1.5, 2, and 2.5).

In general, for the larger population sizes the maximum $|BCB[\cdot]|$ occurs when α and β have their maximum values (upper-left part of the heat map where $\alpha = 0.25$ and $\beta = 2.5$); and the minimum $|BCB[\cdot]|$ occurs in the center of the heat map (where $\alpha = 0.15$ and $\beta = 1$ or 1.5). The maximum $|B[\alpha]|$ occurs in the upper-left part of the heat map for all population sizes and increases with population size. For a population size of 9, the maximum $|B[\alpha]|$ is 0.33 at $\alpha = 0.25$ and $\beta = 1$; for a population size of 81, the maximum is 0.12 at $\alpha = 0.25$ and $\beta = 0.5$. For parameter β, maximum $|B[\beta]|$ occurs when α and β have their minimum values, $\beta = 0.5$ and α =0.05, and the value increases with population size. The maximum $|B[\beta]|$ is 1.78 (when $\alpha = 0.2$ and $\beta = 0.5$) for a population size of 9 and is 0.26 (when $\alpha = 0.05$ and $\beta = 0.5$) for a population size of 81.

Generally speaking, when β has its largest value, $\beta = 2.5$, $|BCB[\beta]|$ performs worse than $|B[\beta]|$ for all population sizes. However, as population size increases the poor performance can also be seen when α has its smallest value (lower-left part of the heat map), $\alpha = 0.05$. In the upper-left part of the heat map, $|BCB[\beta]|$ performs better than $|B[\beta]|$ for all population sizes.

$|BCB[\cdot]|$'s tend to be larger than $|B[\cdot]|$'s for large population sizes (i.e., population sizes of 49 and 81) for all values of α, where β has the highest value (β=2.5). This is also the case when α has its minimum value ($\alpha = 0.05$) .

4.3 Extra Scenarios for BCB in ILMs

The poor performance of the bias correction method in some parts of the parameter space, in some scenarios, may be due to the choice of ILM parameter values defining the training set. We therefore consider two other scenarios with different values for the parameters. In the first scenario, we define our first grid

with α values from 0.03 to 0.07 with a step size of 0.01 and β values from 2.3 to 2.7 with a step size of 0.1. In second scenario, the second grid is defined as α values are from 0.18 to 0.22 with step-size 0.01 and β values are from 0.3 to 0.7 with step-size 0.1. We calculated and compared $|B[\cdot]|$ and $|BCB[\cdot]|$ for each parameter and for each grid.

4.3.1 Scenario 1: Fine resolution grid centered on (α, β)=(0.05, 2.5)

In this scenario, we consider bias correction when the regression design matrix is based on a fine grid with α values from 0.03 to 0.07 with a step-size of 0.01 and β from 2.3 to 2.7 with a step-size of 0.1. In Appendix A, Figures 9a, 10a, 11a, and 12a and 9b, 10b, 11b, and 12b show the absolute mean posterior bias $|B[\cdot]|$ and mean bias corrected $|BCB[\cdot]|$ for parameters α and β, respectively, for the four population sizes. These figures illustrate once again that posterior bias for α and β tends to increase as the population size decreases.

4.3.2 Scenario 2: Fine resolution grid centered on (α, β)=(0.25, 0.5)

In this scenario, we consider bias correction when the regression design matrix is based on a fine grid with α values from 0.23 to 0.27 with a step-size of 0.01 and β from 0.3 to 0.7 with a step-size of 0.1. In Appendix B, Figures 13a, 14a, 15a, and 16a, and Figures 13b, 14b, 15b, and 16b show the absolute mean posterior bias $|B[\cdot]|$ and mean bias corrected $|BCB[\cdot]|$ for parameters α and β, respectively, for the four population sizes. These results show once again that posterior bias for α and β tends to increase as the population size decreases.

4.4 Results from Extra Simulation Scenarios

The figures for Scenario 1 in Appendix A show that using a fine grid centered on (α,β)=(0.05, 2.5) to construct our bias-corrected estimates does not generally improve the bias estimation performance. Although $|BCB[\beta]|$ is almost always smaller than $|B[\beta]|$ under the new fine grid, $|BCB[\alpha]|$ is almost always larger.

The figures for Scenario 2 in Appendix B show the same results as the figures for Scenario 1 in Appendix A. Both results seem to add weight to our earlier conclusion that bias correction in relatively low-bias situations can lead to worse estimates.

5 Summary and Future Work

This paper explore the effects of population size and model parameters on parameter estimation in ILMs. In both homogeneous and spatial models, larger population sizes lead to less bias in parameter estimates.

Probably for this reason, it can be inferred that population size can have a great impact on the performance of a regression-based bias correction method. For large population sizes (i.e., n = 49 or 81), our

bias correction method does not necessarily provide a better estimation for parameters compared with posterior bias for standard posterior mean estimates. However, under such scenarios posterior bias appears low. Overall, these results imply that our bias correction method probably should not be used in general for all population sizes and when the posterior bias is small. Thus, care should be taken to explore the extent of bias that exists (via simulation) before we attempt to correct it.

In this study, we did not considered uncertainty associated with our estimates. Efron and Tibshirani (1993) suggest that standard errors of standard and bias-corrected estimates should be considered. In the context of non-parametric bootstrap, this can lead to highly computationally intensive analyses. In our case, the problem is even worse, as each of our simulations (parametric bootstrap realizations) is analyzed using MCMC in a Bayesian framework. However, this is work that will be of interest in the future. It may be more sensible to use optimization techniques to find maximum posterior (or likelihood) estimates.

In terms of further work, a number of other avenues could be pursued. The goal of this paper was to investigate bias and bias-corrected estimates in spatial and homogeneous SI models.This of course leaves room for investigating SIR, SEIR, or other compartmental frameworks, perhaps with unknown infectious/ latent periods.

Checking bias for network-based models, where networks describe the mixing structure of a population, or for populations with a more heterogeneous spatial structure and/or covariates (e.g., binary covariates representing, say, vaccination status or continuous covariates representing, say, size of farm), would be more realistic. However, in each of these scenarios we would have more parameters to estimate, likely increasing the computational complexity. In addition, exploring the question of whether correcting parameter estimates for bias helps to improve the predictive (e.g., forecasting) qualities of the model is another area for further work. Furthermore, because our bias correction method only tells us how to change point estimates of model parameters, it would be interesting to investigate the estimate of credible intervals to capture parameter uncertainty in bias-corrected estimates.

Another avenue we can consider is testing and improving the goodness of fit of the bias connection model. In our case, we could investigate how well our quadratic model can explain bias in ILMs, or if the use of smoothing splines, or perhaps machine learning algorithms to build models for the bias, might be better than other models for bias correction. Finally, rather than using a series of univariate regression models to model the marginal bias for each parameter, we may find multivariate regression models that account for correlation between parameters could improve our bias-correction method.

References

Almutiry, W., Warriyar, V., and Deardon, R. (2018). Continuous time distance-based and network-based individual level models for epidemics: EpiILMCT. *Journal of Statistical Software*.

Bifolchi, N., Deardon, R., and Feng, Z. (2013). Spatial approximations of network-based individual level infectious disease models. *Spatial and Spatio-Temporal Epidemiology*, 6:59–70.

Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., Swerdlow, D., and the Pennsylvania H1N1 working group (2011). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *PNAS*, 108:2825–2830.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. J. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, 20(1):239–261.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, chapter 10, pages 138–139. Chapman and Hall/CRC.

Gamerman, D. and Lopes (2006). *Markov Chain Monte Carlo Stochastic: Simulation for Bayesian Inference*. Chapman and Hall/CRC.

Kwong, G. P. and Deardon, R. (2012). Linearized forms of individual-level models for large-scale spatial infectious disease systems. *Bulletin of Mathematical Biology*, 74.

Malik, R., Deardon, R., Kwong, G. P., and Cowling, B. J. (2014). Individual-level modeling of the spread of influenza within households. *Journal of Applied Statistics*, 41(7).

O'Neill, P. D. (2010). Introduction and snapshot review: Relating infectious disease transmission models to data. *Statistica Sinica*, 29:2069–2077.

Riley, S., Eames, K., Isham, V., Mollison, D., and Trapman, P. (2014). Five challenges for spatial epidemic models. *Epidemics*, 10:68–81.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E. J., Brooks, S. P., and Grenfell, B. T. (2006). Effect of data quality on estimates of farm infectiousness trends in the UK 2001 foot-and-mouth disease epidemic. *Journal of the Royal Society Interface*, 4:235–241.

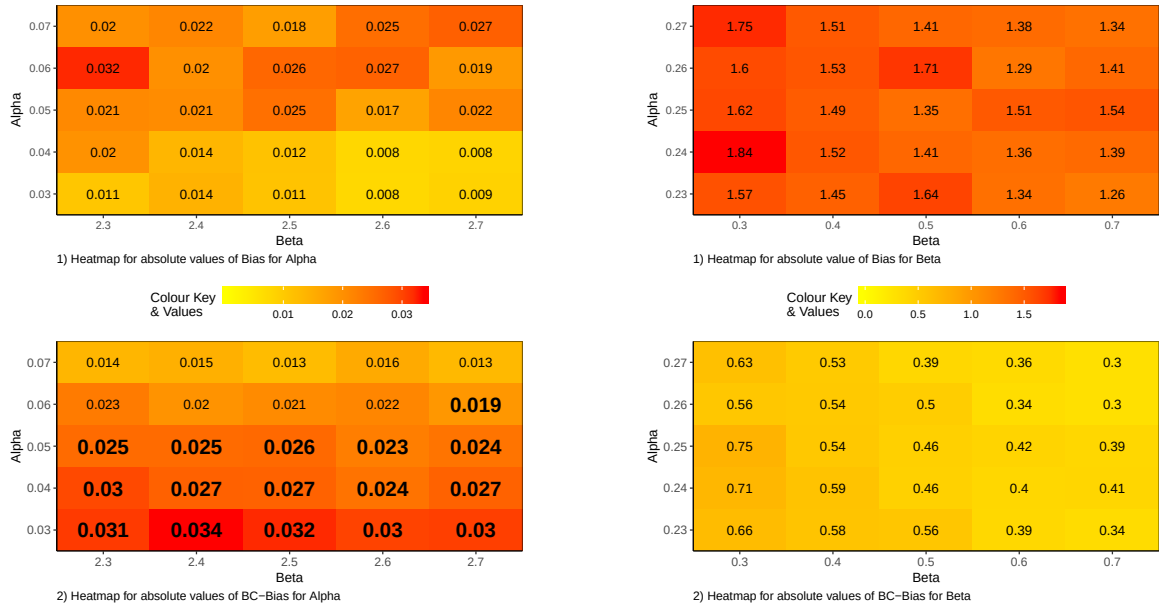
Siettos, C. I. and Russo, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence*, 4:295–306.

Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E. J., Grenfell, B. T., and Keeling, M. J. (2006). Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature*, 44:83–86.

Warriyar, V. and Deardon, R. (2018). Spatial and network based individual level models for epidemics: EpiILM. *Comprehensive R Archive Network*.

Appendix A

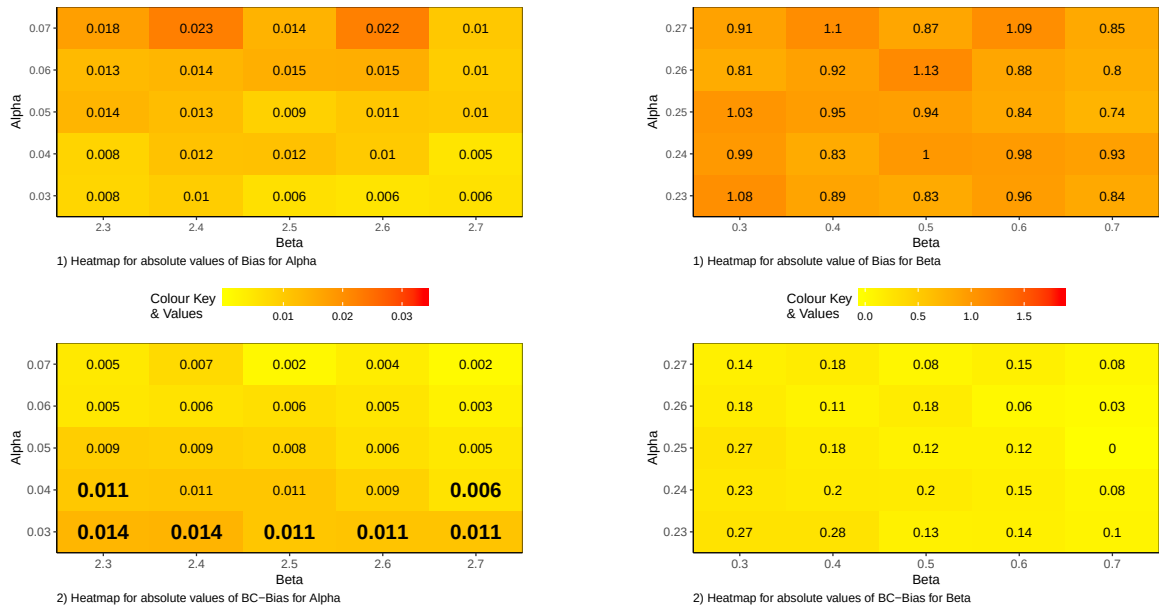
A.1 Fine Resolution Grid Centered on (α, β)=(0.05, 2.5)



(a) Heat maps for absolute values of B and BCB for α

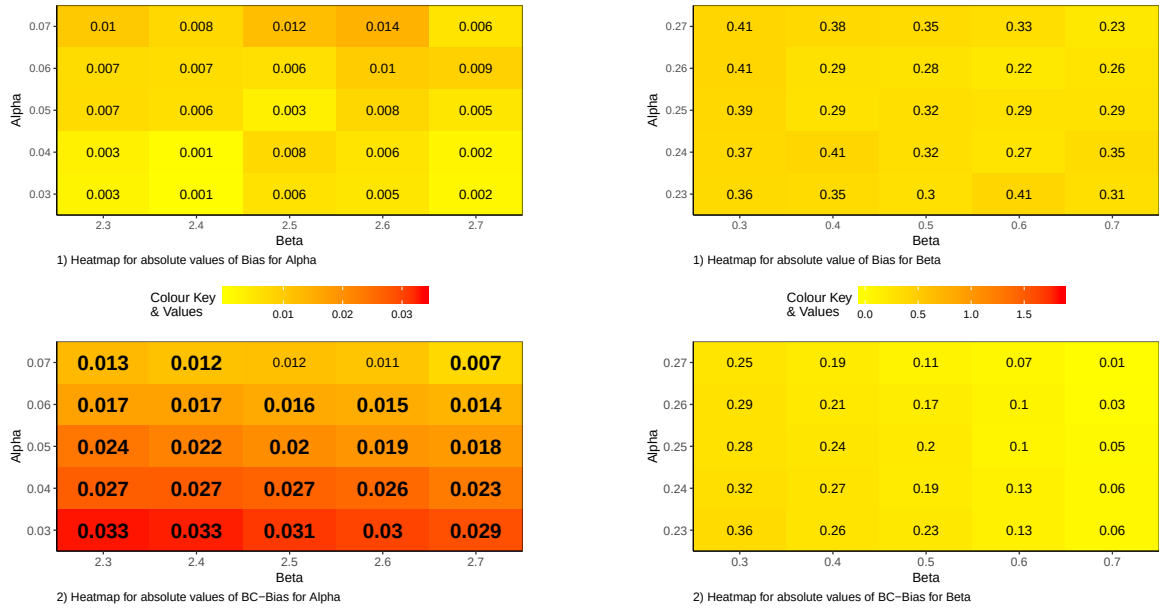
(b) Heat maps for mean absolute values of B and BCB for β

Figure 9: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the first grid (population size n = 9)



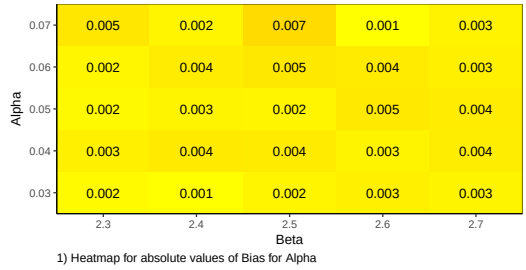
(a) Heat maps for absolute values of B and BCB for α (b) Heat maps for mean absolute values of B and BCB for β

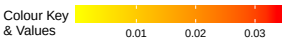
Figure 10: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the first grid (population size n = 16)

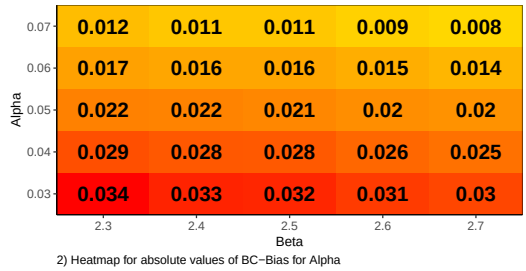


(a) Heat maps for absolute values of B and BCB for α (b) Heat maps for mean absolute values of B and BCB for β

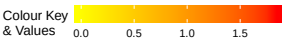
Figure 11: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the first grid (population size n = 49)

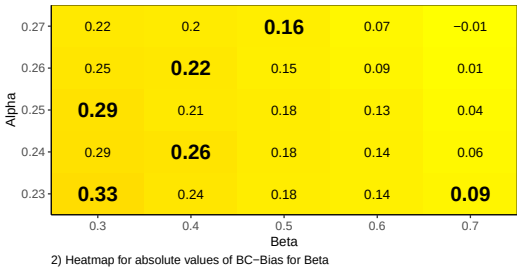






(a) Heat maps for absolute values of B and BCB for α



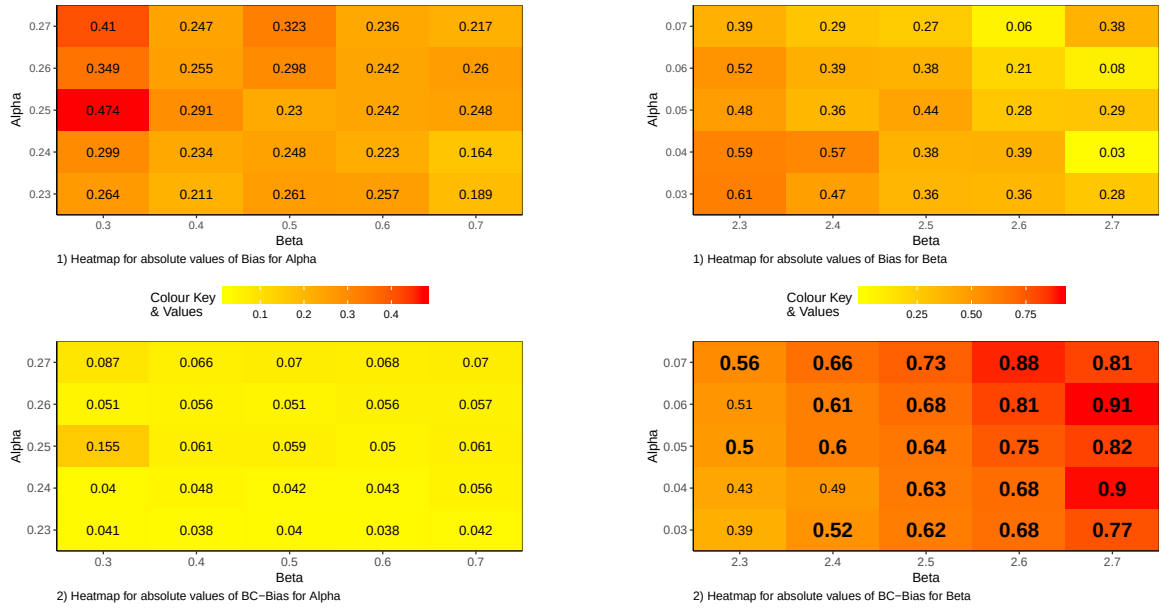


(b) Heat maps for mean absolute values of B and BCB for β

Figure 12: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the first grid (population size n = 81)

Appendix B

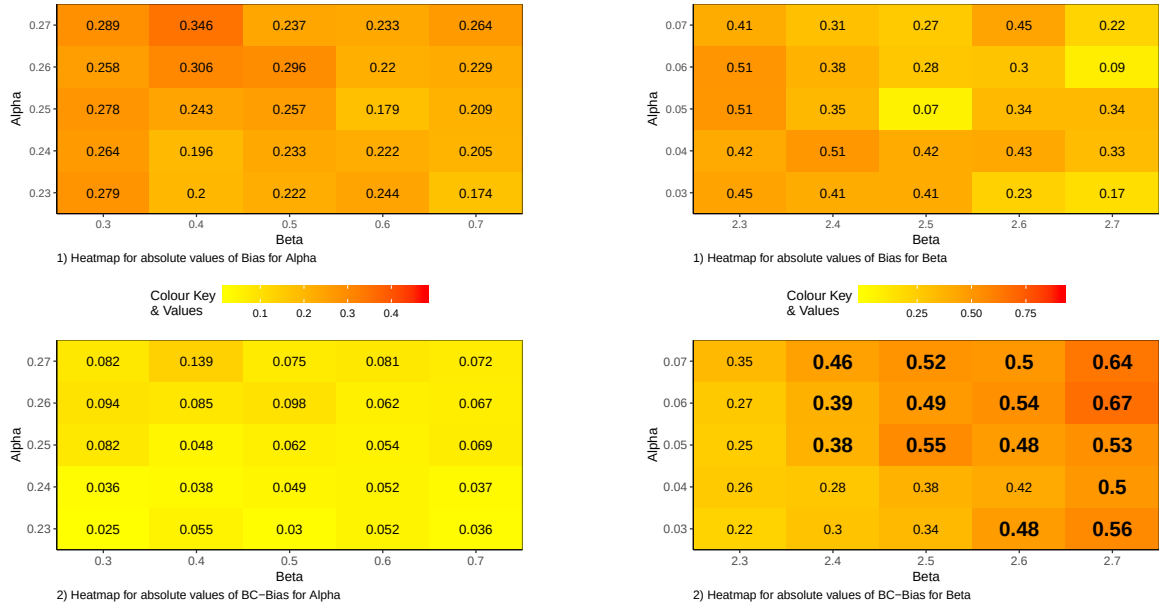
B.1 Scenario 2: Fine Resolution Grid Centered on (α, β)=(0.25, 0.5)



(a) Heat maps for absolute values of B and BCB for α

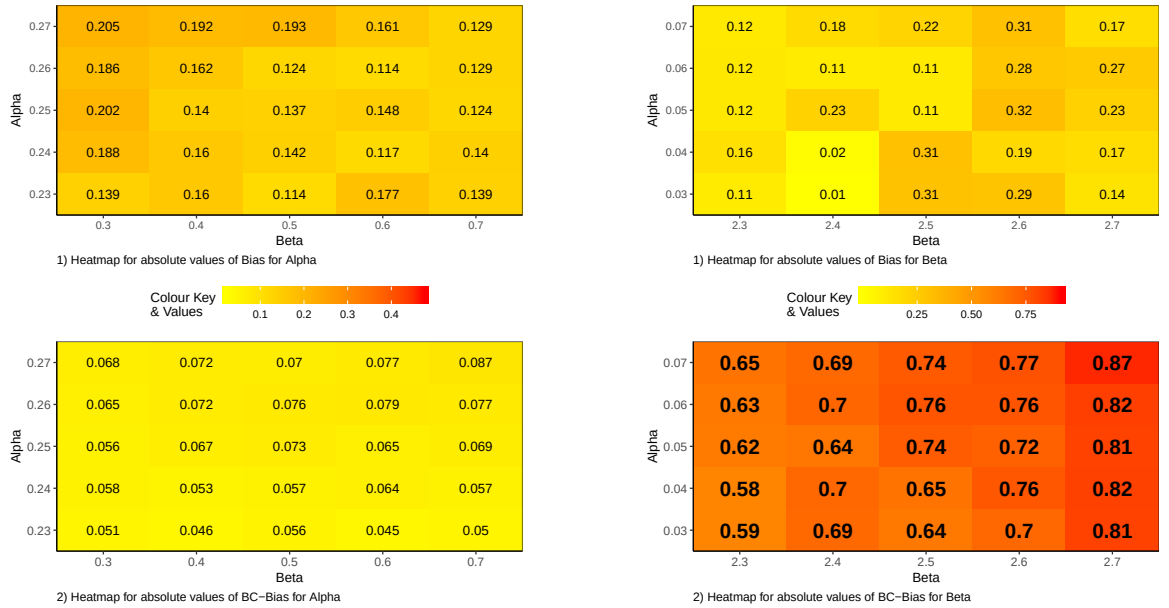
(b) Heat maps for mean absolute values of B and BCB for β

Figure 13: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the second grid (population size n = 9)



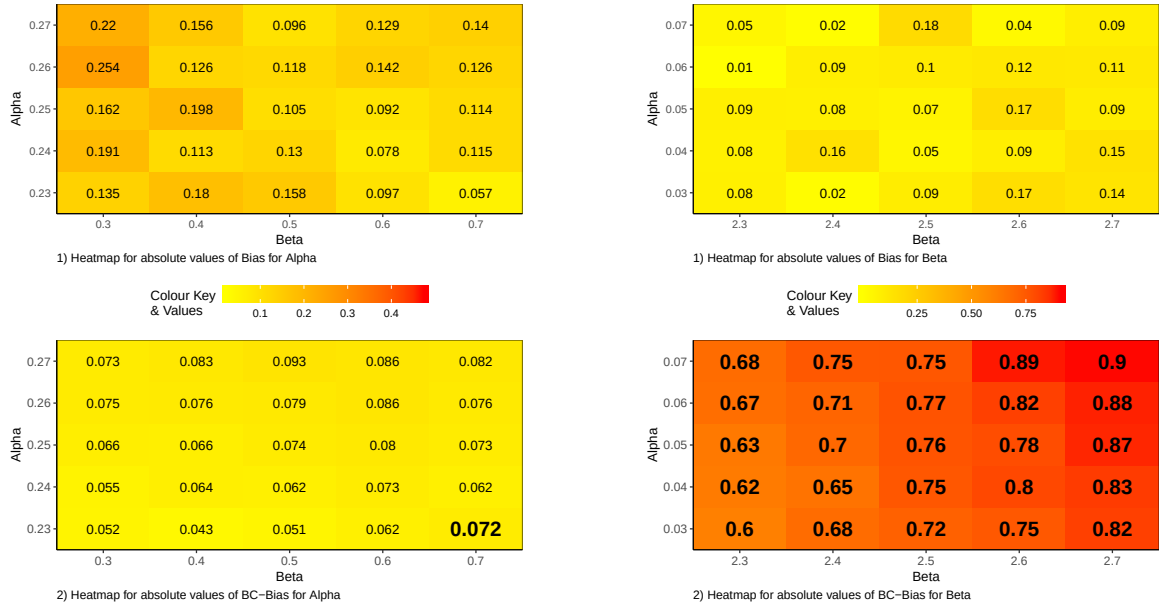
(a) Heat maps for absolute values of B and BCB for α (b) Heat maps for mean absolute values of B and BCB for β

Figure 14: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the second grid (population size n = 16)



(a) Heat maps for absolute values of B and BCB for α (b) Heat maps for mean absolute values of B and BCB for β

Figure 15: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the second grid (population size n = 49)



(a) Heat maps for absolute values of B and BCB for α

(b) Heat maps for mean absolute values of B and BCB for β

Figure 16: Comparison between mean absolute value of B and BCB based on different values of parameters α and β for the second grid (population size n = 81)