Calibrating Models to Data: A Comparison of Methods

by

Zenabu Suboi

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Mathematics in the Faculty of Science at Stellenbosch University

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: 1. Dr. Marijn C. Hazelbag

2. Prof. Wim Delva

December 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: .....
Zenabu Suboi

Date: April 2, 2020 .....

Abstract

Calibrating Models to Data: A Comparison of Methods

Zenabu Suboi

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc. (Mathematics)

December 2020

ddddddddddddddddddddddddddddddddddddddd.

Keywords:

.....

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations ...

Dedications

T.....

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Most sciences today use mathematical and computer simulation models to approximate the real-world processes under study [2, 4, 11]. For example, models play a significant role in health policymaking by estimating the impact of interventions in situations where empirical studies may be time-consuming, costly and impractical [8]. Developing a model calls for a trade-off between computational cost and accuracy; simple models require little computation time but can be a poor description of the real-world process, whereas complex models allow for a more accurate description of the process at the cost of increased computational cost. After model development, it is imperative to know how well the model represents reality. Model calibration, or fitting the model to data, increases the confidence that the model provides a realistic approximation to the real-world process [8, 11] but this process involves running the model many times. This becomes a problem with complex models for which model run-times are very long (Hladish).

Calibration is the process of comparing model outputs with empirical data to identify the model parameter values that achieve a good fit to data [6, 11]. Researchers commonly use calibration methods to find parameter values in case parameter estimates are not available in the literature [10]. The main components of calibration are; summary statistics (targets), the parameter-search strategy, the goodness-of-fit (GOF) measure, acceptance criteria and stopping rules. Improving the computational cost of the calibration involves specification of a more efficient parameter search strategy. In this study, we focus on sampling algorithms as the parameter search strategy, since sampling methods obtain valid

estimates of parameter uncertainty and correlations between parameters. Several more efficient sampling algorithms have been proposed and the number of studies that apply these algorithms is proliferating [11].

Existing literature compares the performance of alternative algorithms for calibrating the same model but does not allow us to draw general conclusions [1, 7]. [3] highlights the need for simulation-based studies that inform the choice of the parameter search strategy in terms of correct estimation of the posterior in different scenarios in terms of contextual variables (i.e. the number of target statistics, the number of calibrated parameters). the performance, strengths and limitations of different model calibration methods. algorithm implementation vs "How the algorithm works on paper"

Because there are many model calibration methods with little or no consensus on their performance, we perform a simulation study to compare the performance of model calibration methods using a simple stochastic Susceptible-Infected-Recovered (SIR) model. The methods to be compared are Rejection Approximate Bayesian Computation (Rejection ABC), Sequential Approximate Bayesian Computation (Sequential ABC) and Bayesian Maximum Likelihood estimation (BMLE).

Outline to be completed when thesis is fully written.....

Chapter 2

Methods

Intro to chapter...

2.1 Calibration Methods Compared

We compare the following calibration methods; Rejection Approximate Bayesian Computation (Rejection ABC), Sequential Approximate Bayesian Computation (Sequential ABC), ... and Bayesian Maximum Likelihood estimation (BMLE). In the next sections, we will shortly explain each of these.

2.1.1 Rejection ABC

Rejection ABC is the most basic form of ABC. This method operates by sampling parameter values $(\theta_i, i = 1, \ldots, N)$ from the prior distribution $\pi(\theta)$ and given these sampled parameter values, data (y) is simulated under a model. A summary statistic (s) of the simulated data (y) must satisfy a proximity criterion with the target statistic (t) of the observed data (x) such that $d(t, s) \leq \epsilon$, where d expresses the distance between the target and summary statistics t and s, and ϵ represents a tolerance level. A target statistic is a data point from the observed data to be considered during the simulation procedure before a decision is made, as to whether a certain parameter combination is to be retained or discarded. The algorithm retains sampled parameter values for which the model produces simulated summaries (s) that are closer to the target statistic (t) than the tolerance (ϵ) [9]. Figure 2.1 illustrates how the rejection ABC al-

gorithm functions. The simulator (M) is run each time with a newly sampled parameter value (θ) from the prior distribution obtaining a simulated summary statistic. When the distance between a summary statistic and the target statistic (y_θ) is smaller than (ϵ); the parameter value is retained (red dot). Otherwise, it is discarded.

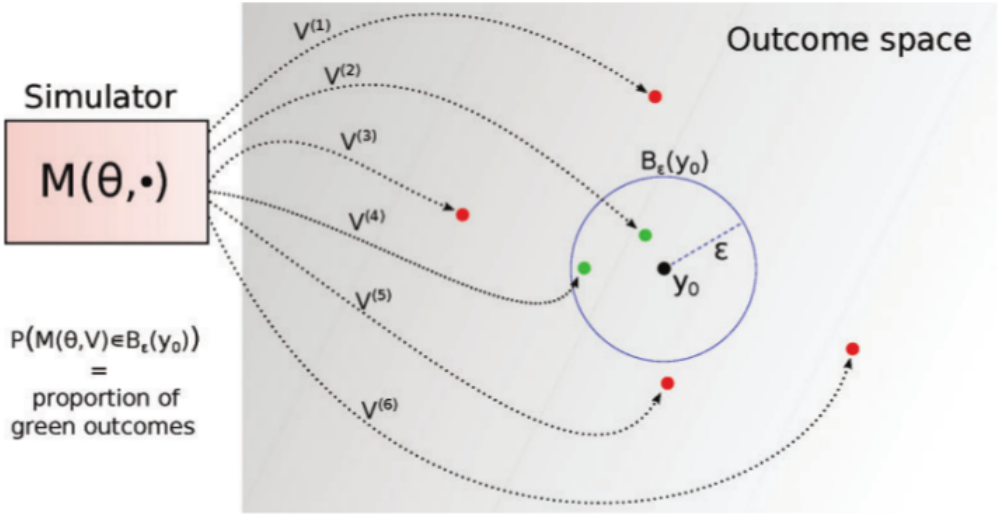

Figure 2.1: Illustration of how the ABC Rejection algorithm works [5].

From the Bayesian framework, estimation of the posterior distribution depends on the prior distribution and the likelihood. The posterior is defined as

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta))}{\pi(x)} \tag{2.1.1}$$

Since the denominator $\pi(\theta)$, which is the marginal probability of the data does not depend on θ, the posterior can be expressed as proportional to the numerator, as shown in equation 2.1.2 below

$$\pi(\theta|x) = \pi(x|\theta)\pi(\theta) \tag{2.1.2}$$

Where θ is a vector of parameter values, x is the observed data, $\pi(\theta|x)$ is the posterior distribution, $\pi(x|\theta)$ is the likelihood and $\pi(\theta)$ is the prior distribution. ABC techniques use this same knowledge in the approximation of the posterior. The distribution of the retained parameter values is expected to converge to the

posterior distribution for arbitrarily small values of the tolerance (ϵ) without the explicit calculation of the likelihood, such that

$$\pi(\theta|x) = \pi(\theta|d(t,s) \leq \epsilon) \tag{2.1.3}$$

2.1.2 Sequential ABC

Sequential ABC is a class of ABC methods that approximates the posterior progressively by drawing samples from the prior sequentially [**?**]. The prior for a particular sampling step depends on the previous retained sample except for the first sampling step which draws from the prior parameter space provided. Thus, the tolerance of the initial sampling step is less restrictive compared to the subsequent ones [**?**]. The sample at the current sampling step $(S^{(t)} = \theta^{(t)}_{(i,i=1,\ldots,N)})$ is derived from the previous sample $(S^{(t-1)})$ using a decreasing sequence of tolerance levels. These methods determine by themselves the tolerance level used at each sampling step and provide a stopping criterion [**?**]. This choice of tolerance for the current sampling step is determined as a function of the distances simulated in the previous sampling step [**?**]. Figure 2.2 gives an illustration of how the sequential ABC algorithm works. The first step of sequential ABC is the same as running rejection ABC; simulator (M) is run with parameter values $(\theta^{(1)})$ sampled from the prior distribution with tolerance (ϵ_1) obtaining a simulated sample $(S^{(1)})$ and retained parameter values $(\theta^{(1)}_r)$. In the second step, tolerance (ϵ_2) is decreased compared to tolerance (ϵ_1) and the prior is determined by the retained parameter combinations $(\theta^{(1)}_r)$ in the first step. A second sample $(S^{(2)})$ of parameter values is obtained at tolerance (ϵ_2) and this process is repeated until a stopping criterion is reached. At each sampling step, a decision is made whether to retain a particular parameter value or discard it. If a simulated summary statistic at that step is further from the target statistic (y_0) than the tolerance level (ϵ_i) of that step, that particular parameter value is discarded, otherwise it is retained. The final sample $(S^{(N)})$aproximates the posterior distribution.

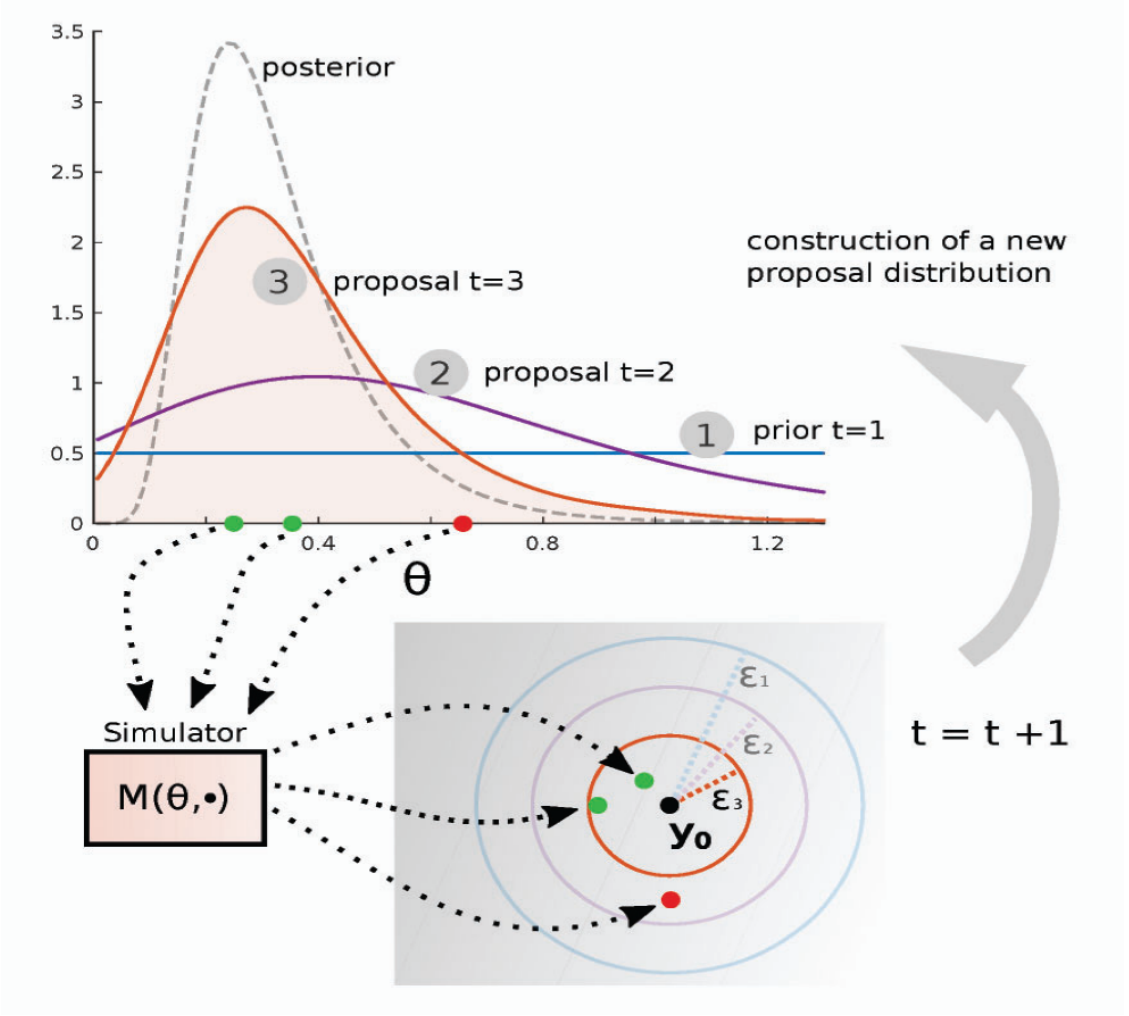

Figure 2.2: Illustration of how the ABC Sequential algorithm works [5].

2.1.3 Bayesian Maximum Likelihood Estimation (BMLE)

Bayesian Maximum Likelihood Estimation (BMLE) termed as "sampling from the posterior distribution" in [6], approximates the posterior by applying sampling importance resampling. The Steps below describe the algorithm of BMLE and how the method is implemented.

- Draw a large number of parameter combinations from the prior distribution
- For each parameter combination, run the model and estimate model outcomes

- Using these model outcomes, estimate the likelihood for the parameter combination and retain this value (log-likelihood)
- Resample from the original parameter sample with replacement, using the likelihood values as sampling weights.

The goodness-of-fit measure used in this model calibration method is the likelihood. Parameter combinations with high values of the likelihood are more consistent with the target supplied. This property allows the assessment of how the data supports one parameter combination compared with another.

2.2 The Simulation Model

2.2.1 The SIR Model

The simulation model used in this study was a simple stochastic SIR (Susceptible - Infectious - Recovered) model, which we used to generate the observed data and applied to the methods. The SIR model is an epidemiological and compartmental model that computes the number of infectious individuals with an infectious disease in a closed population over time. A closed population implies that the population size remains constant over time, that is, there are no births and deaths. The population is divided into three compartments (i.e. health states) – susceptible, Infectious and recovered. The rates in between the compartments determine how many individuals move from one compartment to another.

Figure 2.3: Structure of the simple SIR model.

This model involves a system of three non-linear ordinary differential equations (ODEs) that relates the number of susceptible $S(t)$, number of infectious $I(t)$, and number of recovered $R(t)$ individuals [**?**]. The following system of ODEs governs the dynamics of the SIR model

$$\frac{dS}{dt} = -\beta SI \tag{2.2.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2.2.2}$$

$$\frac{dR}{dt} = \gamma I \tag{2.2.3}$$

Where $\beta > 0$ is the disease transmission rate, $\gamma > 0$ is the recovery rate, $D_{inf} = \frac{1}{\gamma}$ is the duration of infection and $R_0 = \beta D_{inf}$ the basic reproductive number. S is the proportion of individuals in the population that are susceptible to the disease and I represents the proportion of infectious individuals. Susceptible individuals become infectious at a rate β. At a rate γ , infectious individuals recover from the disease (gain permanent immunity to the disease) (LStone). The population stays constant throughout the transmission dynamics over the set time such that

$$S(t) + I(t) + R(t) = N \tag{2.2.4}$$

Figure 2.3 illustrates the dynamics of a stochastic SIR model run in the R software over time (t) of 75 days for a population (N) of 1000 individuals. The blue curve indicates the Susceptible compartment, the red curve indicates the Infectious compartment and the green curve indicates the Recovered individuals. The susceptible compartment reduces to zero as the infectious compartment gradually picks up.

2.3 Performance Measures

To compare the performance of the methods, we ran equal number of simulations for all methods per scenario and compared the resulting posterior to the reference posterior by computing percentage overlaps.

2.3.1 Recording Efficiency (Computational Cost)

To record how efficient each method is, we ran equal number of simulations for each model calibration method. For each scenario and calibration method, we recorded the total time taken to perform simulations and the time taken to

run the SIR model in each calibration method. From these recorded times, we obtained the algorithm implementation times for each method as follows:

$$\text{Algorithm implementation time} = \text{Total runtime} - \text{Model runtime} \tag{2.3.1}$$

2.3.2 Percentage Overlap

To compare the posterior densities of the methods to the reference posterior density, we created a raster using the raster function from the raster library in R, which we used to compute percentage overlaps. A raster consists of a matrix of cells or pixels arranged into rows and columns to form a grid. Each cell contains a value which is the number of observations counted within a particular cell and represented by a color gradient. The raster was created by considering the minimum and maximum values of beta and gamma retained by the calibration methods to be compared as well as the reference. This was done so that the same raster could be applied to all the methods and reference. The resulting parameter space was divided into 100 x 100 equally sized bins with beta values on the x-axis and gamma values on the y-axis (see Figure 2.4). This formed a grid in which the posterior densities laid. We applied the grid to each posterior density to quantify the density of each cell or pixel.

The percentage overlap for each method was computed by summing the within cell density differences between the calibration method and the reference for that particular scenario and subtracting from 1. This was done such a way that, in the case of a perfect overlap between calibration method and the reference, percentage overlap goes to 1 and in the case where there is no overlap between calibration method and the reference, percentage overlap goes to 0 (see equation 2.3.2 below).

$$P_{ij} = [1 - \frac{(\sum |M_{ij} - R_i|)}{2n}]x100 \tag{2.3.2}$$

Where P_{ij} is the percentage overlap for method j and scenario i, M_{ij} represents the matrix form of the raster applied on calibration method j and scenario i, R_i represents the matrix form of the raster applied on the reference for scenario i, n is the number of parameter combinations retained by each calibration method.

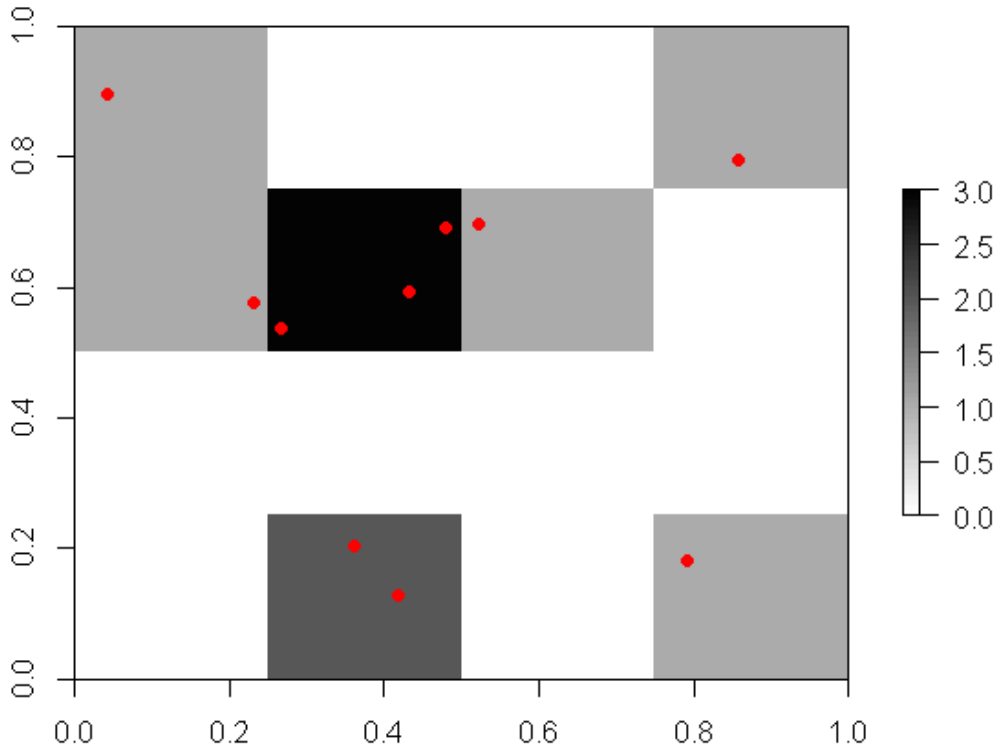

Figure 2.4: 4x4 raster applied to a posterior density.

Chapter 3

Simulation Procedure

A simulation study was conducted which allows changing parameters of interest (i.e. the number of target statistics and number of simulations). All simulations and analyses were conducted using R version 3.6.2 (2019-12-12)

3.1 Simulation Setup

3.2 Obtaining Targtes

As target statistics in this study, there were two scenarios. Scenario one (S1) considered two target features (prevalence at two-time points, 50 and 75) and scenario two (S2) looked at three target features (peak prevalence for each model run in addition to the two-time points in scenario one).

R version 3.6.2 (2019-12-12) was used to perform the statistical analyses and datasets were obtained from a stochastic SIR model using the SIR function in the SimInf library [**?**]. To obtain targets for scenario one, the SIR model was ran once with known parameter values $\beta = 0.2, \gamma = 0.02$ and the prevalence at times 50 and 75 were saved as target statistics. Similarly for scenario two, the SIR model was ran once with the same known parameter values and the prevalence at times 50 and 75, as well as the peak prevalence, were saved as targets (see Figure **??** below).

Table 3.1: Simulation Setup for Calibration Methods

	Scenario One	Scenario Two
Targets features	Prevalence at times 50 and 75	Prevalence at times 50 and 75 + peak prevalence
Targets used	Prev. at 50 = 0.644 (644 out of 1000) Prev. at 75 = 0.404 (404 out of 1000)	Prev. at 50 = 0.622 (622 out of 1000) Prev. at 75 = 0.371 (371 out of 1000) Peak prev. = 0.677 (677 out of 1000)
Number of simulations	60,000	75,000
arameters P used to generate targets	Beta = 0.2 Gamma = 0.02	Beta = 0.2 Gamma = 0.02
nitial compIartmental values	S=990 I=10 R=0	S=990 I=10 R=0
Time span	Sequence from 0 to75 incremented by 1	Sequence from 0 to75 incremented by 1
Beta prior	Uniform distribution (0,1)	Uniform distribution (0,1)
Gamma prior	Uniform distribution (0,0.5)	Uniform distribution (0,0.5)

Table 3.2: Simulation Setup for the Reference

	Scenario One	Scenario Two
Targets features	Prevalence at times 50 and 75	Prevalence at times 50 and 75 + peak prevalence
Targets used	Prev. at 50 = 0.644 (644 out of 1000) Prev. at 75 = 0.404 (404 out of 1000)	Prev. at 50 = 0.622 (622 out of 1000) Prev. at 75 = 0.371 (371 out of 1000) Peak prev. = 0.677 (677 out of 1000)
Number of simulations	1,000,000	1,000,000
arameters P used to generate targets	Beta = 0.2 Gamma = 0.02	Beta = 0.2 Gamma = 0.02
Beta prior	Uniform distribution (0.1,0.4)	Uniform distribution (0.1,0.4)
Gamma prior	Uniform distribution (0.01,0.03)	Uniform distribution (0.01,0.03)

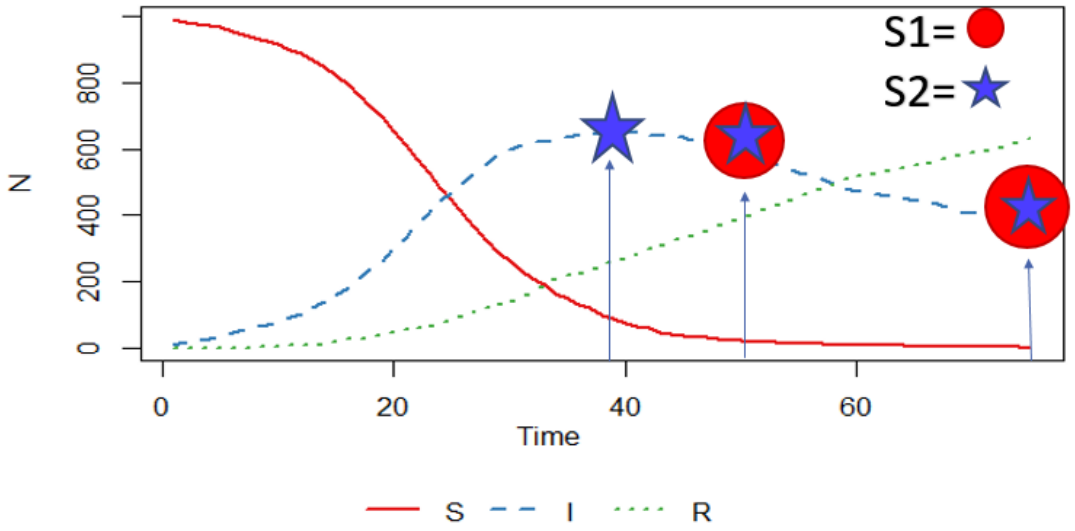

Figure 3.1: targets for scenarios 1 and 2. Scenario 1 considers only green points on the Infectious (I) curve while scenario 2 considers both red and green points on the (I) curve

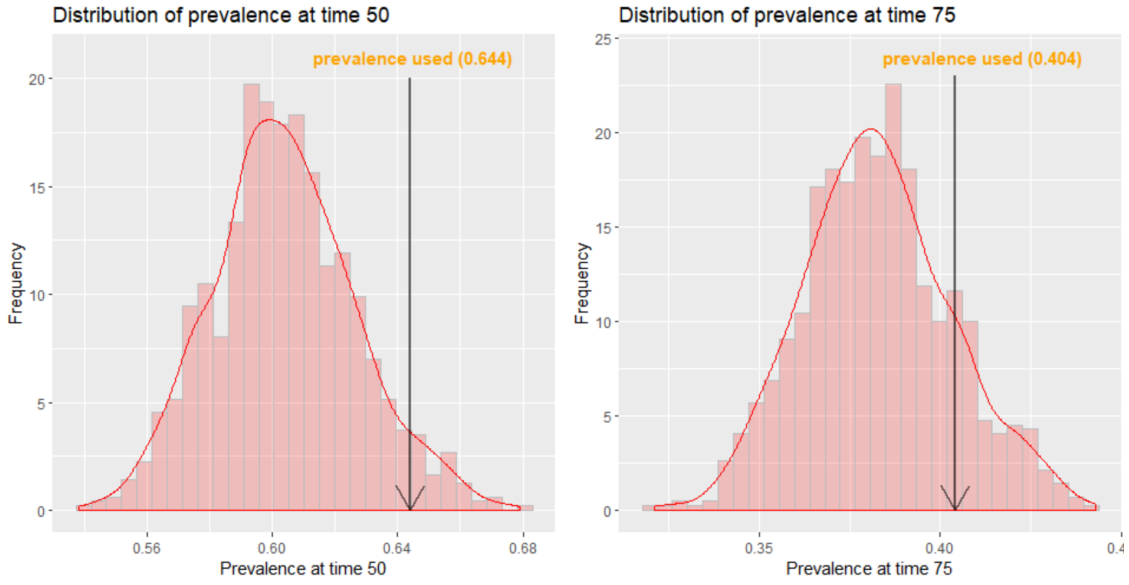

Figure 3.2: Distribution of targets in Scenario 1

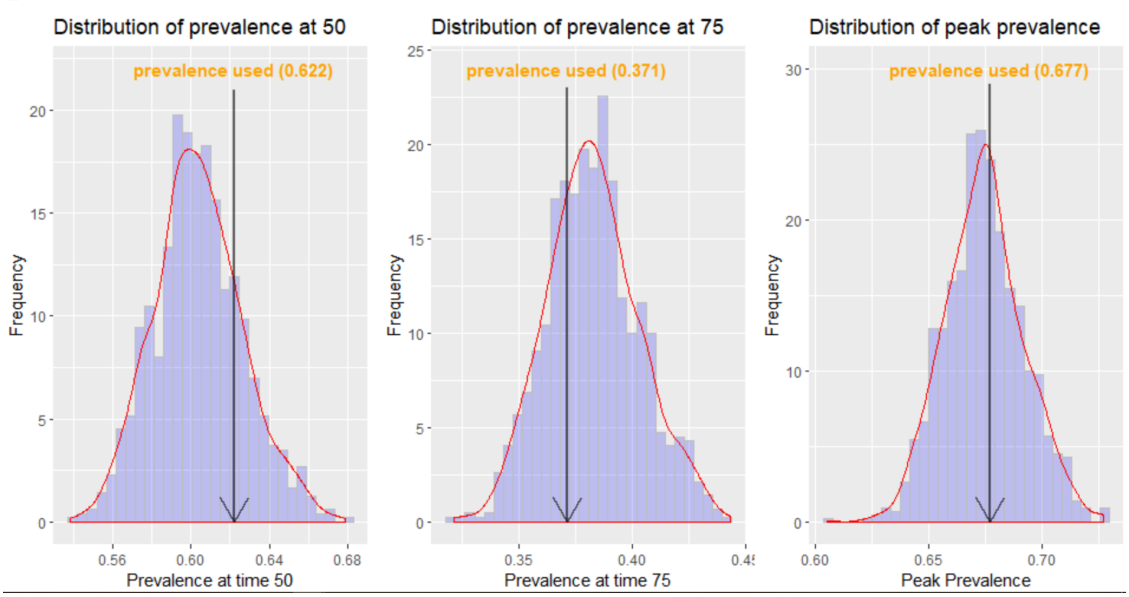

Figure 3.3: Distribution of targets in Scenario 2

Chapter 4

Results

Chapter 5

Discussion and Conclusion

5.1 Discussion

5.2 Conclusion

List of references

[1] Issa J Dahabreh, Jeffrey A Chan, Amy Earley, Denish Moorthy, Esther E Avendano, Thomas A Trikalinos, Ethan M Balk, and John B Wong. A review of validation and calibration methods for health care modeling and simulation. In *Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment [Internet]*. Agency for Healthcare Research and Quality (US), 2017.

[2] Anthony T Fojo, Emily A Kendall, Parastu Kasaie, Sourya Shrestha, Thomas A Louis, and David W Dowdy. Mathematical modeling of "chronic" infectious diseases: unpacking the black box. In *Open forum infectious diseases*, volume 4, page ofx172. Oxford University Press US, 2017.

[3] C Marijn Hazelbag, Jonathan Dushoff, Emanuel M Dominic, Zinhle E Mthombothi, and Wim Delva. Calibration of individual-based models to epidemiological data: a systematic review. *medRxiv*, page 19006056, 2019.

[4] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[5] Jarno Lintusaari, Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, 66(1):e66–e82, 2017.

[6] Nicolas A Menzies, Djøra I Soeteman, Ankur Pandya, and Jane J Kim. Bayesian methods for calibrating health policy models: a tutorial. *PharmacoEconomics*, 35(6):613–624, 2017.

[7] Amanda Minter and Renata Retkute. Approximate bayesian computation for infectious disease modelling. *Epidemics*, 29:100368, 2019.

[8] Natasha K Stout, Amy B Knudsen, Chung Yin Kong, Pamela M McMahon, and G Scott Gazelle. Calibration methods used in cancer simulation models and suggested reporting guidelines. *Pharmacoeconomics*, 27(7):533–545, 2009.

[9] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.

[10] Elske van der Vaart, Mark A Beaumont, Alice SA Johnston, and Richard M Sibly. Calibration and evaluation of individual-based models using approximate bayesian computation. *Ecological Modelling*, 312:182–190, 2015.

[11] Tazio Vanni, Jonathan Karnon, Jason Madan, Richard G White, W John Edmunds, Anna M Foss, and Rosa Legood. Calibrating models in economic evaluation. *Pharmacoeconomics*, 29(1):35–49, 2011.