Effective Population Size in Infectious Disease Models

by

Madi Yerlanov

B.Sc., Nazarbayev University, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Mathematics
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	**Madi Yerlanov**
Degree:	**Master of Science**
Thesis title:	**Effective Population Size in Infectious Disease Models**
Committee:	**Chair: Cedric Chauve** Professor, Mathematics
	Caroline Colijn Supervisor Professor, Mathematics
	Nathan Ilten Committee member Associate Professor, Mathematics
	Paul Tupper Examiner Professor, Mathematics

Abstract

The effective population size N_e was introduced by geneticist Sewall Wright to describe idealized populations. N_e has been a research interest because of its mathematical theory and population management utility. Inspired by such potential, we (re)-introduce the notion of the effective population size N^* in mathematical epidemiology. Our aim is to see if a simple model with the population size as a variable N^* can capture disease dynamics in various data types. We introduce a simple SIR model and derive methods of estimating N^*. We apply our methods to both simulated and real outbreak data. We compare N^* to N and look at how corresponding solution curves match data. We identify preferable methods and settings where these methods are applicable. We state possible implementations of N^* in public health management as well as extensions and limitations of our methodology.

Keywords: effective population size, epidemiology, infectious disease modelling, parameter estimation, COVID-19

Dedication

I dedicate this work to Him, Her and Them.

Acknowledgements

I want to thank my supervisor, Caroline Colijn. Caroline saw something in me. I am honored to be her graduate student. She inspired me and I changed my field completely. Her dedication to her students led me to where I am.
I want to thank my mentor, Jessica Stockdale. Each week, she gave me a detailed review of what I wrote. Without her commitment educating me on how to write a research piece, I would not be where I am.
I want to thank the department: staff, professors and students. I had a great time learning new things in mathematics, live in Canada and many others. Each of my classes has that one assignment that I will never forget. Shout out to Department teas and Graduate support sessions. This community made me come to where I am.
I want to thank my family, friends and anyone who made me smile at some point. I had a great time with a lot of my peers even in this pandemic situation. Support from close ones let me get over negative things and reach where I am.
Where I am is the happy place.

Table of Contents

Declaration of Committee ii

Abstract iii

Dedication iv

Acknowledgements v

Table of Contents vi

List of Tables viii

List of Figures ix

1 Preliminaries 1
- 1.1 Introduction . . . 1
- 1.2 Background . . . 2
  - 1.2.1 Effective population size in genomics . . . 2
  - 1.2.2 Compartmental models in epidemiology . . . 4
  - 1.2.3 Effective population size in epidemiology . . . 6

2 Computation of the effective population size 8
- 2.1 Method discussion and derivation . . . 8
- 2.2 Verifying via simulations . . . 12

3 Application to outbreaks of COVID-19 in Chinese cities 23
- 3.1 Data selection . . . 23
- 3.2 Application of methods to estimate the effective population size . . . 28
- 3.3 Discussion on results of methods applied to real data . . . 36

4 Complex model 40
- 4.1 Patch models . . . 40
- 4.2 Complex model simulation . . . 41
- 4.3 Complex model results . . . 43

5 Conclusions **51**
5.1 Final discussion 51
5.2 Summary and Future work 53

Bibliography **56**

Appendix A Supplemental figures **60**

List of Tables

Table 2.1 The methods that will be used throughout this thesis for estimation of N^*. 12

Table 3.1 AICc computed for 3 methods that provide best fitting for chosen 53 cities. Colors denotes the order of scores in each city: highest, middle, lowest. 34

Table 4.1 Transmission rate matrices $\{\beta_{j,i}\}$ for complex model simulation on 3 patches. Short names and number labels are provided under each. All the values get multiplied by 0.0001. 43

List of Figures

Figure 1.1 Early COVID-19 outbreak in Tianjin with population of 15.6 million. The orange dots represent I-data of the city. The blue line represents the fitted I-component of the standard model with the given population size. $\beta = 2.12 \times 10^{-8}$, $\gamma = 0.28$. 7

Figure 2.1 Box-plots for values of N^* obtained from the discussed methods applied to simulated outbreaks data (short names are used). The blue line denotes the true value N. The red dot denotes the mean of 1000 simulations. "Larger" or "smaller" are with respect to the baseline setting parameter. 15

Figure 2.2 Fitting to I-data applied to 1000 simulations in the baseline setting: $N = 1000, \beta = 0.0005, \gamma = 0.2$. Curves are either simulated data '-data' or corresponding deterministic solutions '-solution' that use parameters computed by the method. Blue denotes outliers (data that provided a parameter value beyond the interval $[q_{0.25} - 1.5 * IQR, q_{0.75} + 1.5 * IQR]$ or a corresponding solution), red denotes non-outliers. The deterministic solution with true parameters is shown by the black curve. 18

Figure 2.3 Examples of partially available simulated data and simulated data under lockdown. $N = 1000$, $\beta = 0.0005$, $\gamma = 0.2$. 20

Figure 2.4 Box-plots for values of N^* (y-axis) obtained from the discussed methods under varying levels of partial or lockdown data. The blue line denotes $N = 1000$. The red dot denotes the mean. 21

Figure 3.1 Location of 53 selected cities in China. Colors stand for the maximum number of currently infected individuals in a day (I_{max}): green 50-99, yellow 100-249, orange 249-999, red 1000+. 25

Figure 3.2 I-data for 53 cities divided into groups based on I_{max}. 26

Figure 3.3 I-data for 53 cities divided into groups based on I_{max}. 27

Figure 3.4 N^* results for cities with $I_{max} \in [1000, \infty)$. 95 % CI. x-axis is on a log scale. Numbers in brackets indicate the census population size in millions. 28

Figure 3.5 N^* results for cities with $I_{max} \in [250, 999]$. 95% CI. x-axis is on a log scale. Numbers in brackets indicate the census population size in millions. 29
Figure 3.6 N^* results for cities with $I_{max} \in [100, 249]$. 95% CI. Numbers in brackets indicate the census population size in millions. Fitting to R results are omitted for Bengbu, Zhumadian, Zhengzhou, Fuyang. . 30
Figure 3.7 N^* results for cities with $I_{max} \in [50, 99]$. 95% CI. Numbers in brackets indicate the census population size in millions. Fitting to R results are omitted for Liuan, Putian, Shangqiu, Zhuhai, Huizhou, Ganzhou. 31
Figure 3.8 Fitting results for Tianjin. Orange dots stand for data. Blue lines stand for corresponding solutions. 32
Figure 3.9 Fitting results for Bingbu. Orange dots stand for data. Blue lines stand for corresponding solutions. 32
Figure 3.10 Fitting results for Yichun. Orange dots stand for data. Blue lines stand for corresponding solutions. 33
Figure 3.11 A heatmap of SSE for fixed γ method, Tianjin. SSE is between I-data and I-solution. γ is fixed at 0.167. The blue point denotes the parameter values computed by this method, $(\beta_{optimum}, N^*_{optimum})$. Parameter ranges are from 0.1 to 5 with respect to optimum parameter values computed by fixed γ method, i.e. $\beta \in [0.1\beta_{optimum}, 10\beta_{optimum}]$, $N^* \in [0.1N^*_{optimum}, 10N^*_{optimum}]$. SSE is on the log scale. The red contour denotes $log(SSE)$ that is 1.01 multiple of $log(SSE)$ produced by the blue point (not present, since there are no parameter combination giving lower SSE). 35
Figure 3.12 A heatmap of SSE for fitting to R method, Tianjin. SSE is between R-data and R-solution. β is fixed at the optimum value computed by this method, $\beta_{optimum}$. The blue point denotes the parameter values computed by this method, $(\gamma_{optimum}, N^*_{optimum})$. Parameter ranges are from 0.1 to 5 with respect to optimum parameter values computed by fixed γ method, i.e. $\gamma \in [0.1\gamma_{optimum}, 10\gamma_{optimum}]$, $N^* \in [0.1N^*_{optimum}, 10N^*_{optimum}]$. SSE is on the log scale. The red contour denotes $log(SSE)$ that is 1.01 multiple of $log(SSE)$ produced by the blue point. 36
Figure 3.13 The relationship between N^*/N and distance between a city and Wuhan. N^* is computed using fixed γ method (optimum). Colors correspond to the size of outbreak reflected in I_{max}: large $I_{max} \in [1000, \infty)$, big $I_{max} \in [250, 999]$, medium $I_{max} \in [100, 249]$, small $I_{max} \in [50, 99]$. 39

Figure 4.1 On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axis is restricted to $[0, 6000]$. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula. 47

Figure 4.2 On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axis is restricted to $[0, 6000]$. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula. 48

Figure 4.3 On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axis is restricted to $[0, 6000]$. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula. 49

Figure 4.4 On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axes are restricted to $[0, 10000]$ and $[0, 18000]$ respectively. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula. 50

Figure 5.1 Summary on methods performance based on their application in chapters 2, 3 and 4. The details (e.g. what data and parameters are needed) on each of the methods can be found in section 2.1 and table 2.1. 52

Chapter 1

Preliminaries

1.1 Introduction

Infectious disease modeling is a subfield of mathematical biology that employs mathematical tools to study disease dynamics. Mathematics helps to estimate key parameters and those parameters in turn help to predict progressions of an epidemic. One of such parameters is the basic reproduction number and it may show conditions for disease eradication. Informed with parameters and outcomes of an emerging disease, public health management teams decide how and what control measures (e.g. lockdowns, vaccinations) should be applied. In addition to forecasting potential, infectious disease modeling may be beneficial in retrospective analysis. It allows us to look for causes and interventions of past outbreaks and to deepen our understanding of epidemiology.

Our goal is to investigate population sizes in disease outbreaks. A population size is a key parameter in modelling epidemics, as it shows the scope of outbreaks and appears in other important estimates such as the basic reproduction number. In general, it is not treated as a variable, but in this thesis, we specifically treat it as unknown. We focus on simple methods or methods coming from standard models to explore reliable ways of finding the number of individuals involved in disease dynamics. The aim is to see if a simple model with an unknown population size can capture disease dynamics in outbreaks data where one would use computationally intensive methods. In addition, the implementation of simple methods would allow conveyance of ideas more easily compared to complex models, even for the general public. As a simple model involves a fewer number of parameters and compartments, a disease transmission picture would be clearer. We introduce and develop a notion of "the effective population size" that is coming from evolutionary theory. In the field of genetics, the effective population size N_e is defined as the number of individuals in an idealized population that would reflect key parameters such as genetic drifts. N_e has been implemented because the census population size is not always proportional to given rates such as variance in gene frequency [23]. Similarly, in early outbreaks, the true population size of a location (e.g. a city, a town, a region etc.) could be considerably larger than

transmission analysis suggests. For example, if an outbreak was localized and/or lockdown was implemented, and consequently, one needs to look and employ a different population size since using the census population size and a simple model will lead to a discrepancy between data and a prediction curve.

In this thesis, we first explore the properties of a simple deterministic model and its fitting to data. We analyse methods to compute a population size from this model. We test these methods on stochastically simulated data with the expectation that the methods would provide an effective population size that is almost equal to the true population size on average since mixing in simulations is homogeneous and unrestricted. We attempt to identify robust and accurate ways to find an effective population size. Second, we apply our theoretical findings to early COVID-19 outbreaks in Chinese cities. In particular, we focus on the effective population size and look at how model captures real outbreak data. Third, we develop a more complex model. The complexity of this model will lie in the various mixing patterns among subpopulations. The complex model is used for simulations and we explore whether methods for finding the effective population size and subsequent simple model fitting are still applicable in this heterogeneous setting. We conclude with a discussion on limitations, extensions and implications of the effective population size and associated methods of computing.

1.2 Background

Both genetics and epidemiology saw major developments in the twentieth century and the connection between these two fields is getting stronger both in theory and application. For example, the rapidly expanding area of phylogenetics helps to shed a light on the origins of infections such as influenza and has allowed for more detailed monitoring of its spread [25]. We state relevant key concepts and review relevant topics from both fields. Then we use those notions as a basis to (re)-introduce a new one — the effective population size in epidemiology.

1.2.1 Effective population size in genomics

Charles Darwin founded modern biology by stating the revolutionary theory on natural selection in the middle of the nineteenth century [10]. In the early twentieth century, the rediscovered work of Gregor Mendel led to the emergence of one of the cores of modern biology - genetics, the study of genes [33]. Genetics helps us to answers questions of heredity and can be useful in various ways, for example, diagnosis of cancer [45]. Ronald A. Fisher and Sewall Wright developed a mathematical foundation of population genetics with one of the focal points being *genetic drift* — the change in gene frequencies, caused by random sampling of individuals [9]. For example, in small populations with alleles (gene variants) of the same proportion (with respect to the number of individuals) not all pre-existing

genotypes might be reproduced in the next generation which leads to genetic drift. The role of genetic drift in evolution has been in debate for a century, as its effect is compared to one of natural selection. Nevertheless, it has continued to be a source for active research, especially, in mathematics [15].

Genetic drift can be described by the Wright-Fisher model in an idealised (conveniently simplified) population. This stochastic model assumes random mating, constant population size and discrete simultaneous generations among other conditions. $2N$ genes replicate themselves (or N diploid individuals mate) following those assumptions. The sampling probability follows a binomial distribution $\mathrm{Bin}(2N, p)$. For no drift to occur, it is expected that there is no change in allele frequency, p,

$$\mathrm{E}(p_i|p_{i-1}) = p_{i-1} = \cdots = p \tag{1.1}$$

where p_k is the frequency of a gene variant at generation k. Most importantly, this model allows us to quantify variance

$$\mathrm{var}(p_i|p_{i-1}) = \frac{p_{i-1}(1-p_{i-1})}{2N}. \tag{1.2}$$

With the observed variance $\widehat{\mathrm{var}}$ of the actual (not idealised) population, one can obtain

$$N_e = \frac{p_{i-1}(1-p_{i-1})}{2\widehat{\mathrm{var}}(p_i|p_{i-1})}. \tag{1.3}$$

The *effective population size* (N_e) is then defined as the number of individuals in the idealized population that has the same characteristics (e.g. variance of frequencies) as in the real population of size N [46]. One of the major properties is that $N_e \leq N$, because of the consequences of simplifying assumptions. For example, not all individuals are mating (thus their genotype is not represented) due to an uneven sex ratio. In addition, since there are other genetic characteristics of a population such as inbreeding probability, and a single parameter cannot summarize them all, there are various ways to define and compute this quantity N_e [15].

Because interpretations, estimations and applications of the effective population size are diverse and open, N_e has been studied extensively. One of the questions is to explore the relationships between ecological or demographic features and the effective population size [15]. These features include age structure (maturation and breeding ages), population structure (subpopulations and isolation) and family size (heritability). N_e combined with ecological factors is used to describe human societies as in the example of the tribe study [47]. In this study, computation of N_e helped to develop understanding of the reproductive mechanism in a tribal isolate and how mating was different from the outside human societies. On the other hand, N_e solely can help in the conservation management of endangered species. For example, if a population with small N_e may have a genetic variation loss, one

needs to look at the causes and construct management to maintain diversity (e.g. habitat relocation and/or protection) [44].

Inspired by utility and potential for research, we attempt to translate this notion into the epidemiological setting.

1.2.2 Compartmental models in epidemiology

Epidemiology in its original sense means "study upon/on people", but its first association with infectious disease appeared only in the early nineteenth century [31]. Nevertheless, the first mathematical studies were conducted even before, for example, the eighteenth century work of Daniel Bernoulli on smallpox [4]. In the twentieth century, the term started expanding to include non-communicable (e.g. stroke) as well as non-human (e.g. blight) diseases. The early twentieth century also marks the foundation of mathematical epidemiology with the work of Ross on malaria that utilized a simple compartmental model and introduced a notion of the basic reproduction number [39] [7]. The subject was further elevated by the works of Kermack and McKendrick with the emergence of the first deterministic compartmental model with dependence on the age of infection [19] [20] [21]. This model serves as a basis of the models utilized in this paper.

The *compartmental model* is a special type of epidemiological model where each individual is in exactly one of the compartments (stages of infection) and able to move between them in time [31]. Systems of ordinary differential equations are often used to describe the dynamics between these classes. One of the simplest models is SIR. Susceptible, Infected, and Removed are the respective compartments, each is a function that shows the number of people (at the respective stage of infection) at time t. The SIR model without demography (no birth/death/migration rates) can be described by the following system of differential equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{1.4}$$

where β is the infection/transmission rate and γ is the recovery/removal rate (both are per capita). Individuals move from S to I with rate βSI (i.e. infection) and move from I to R with rate γI (i.e. removal). The population size is $N := S + I + R$, and it is unchanging in time due to earlier assumptions on demography. If one considers outbreaks of a short period, this assumption can be valid.

A central quantity in epidemiology is $\mathcal{R}_0$, the *basic reproduction number*. $\mathcal{R}_0$ is the number of individuals that will be infected by a single infective in a fully susceptible population [31]. One of the important aspects of this quantity is its threshold: if $\mathcal{R}_0 > 1$ disease persists

and if $\mathcal{R}_0 < 1$ disease is eradicated. Note that there are various definitions and methods to compute it such as a spectral radius of the associated next-generation matrix. It cannot always serve as a cut-off, for example, in cases of reinfection (e.g. individuals move from R to S). Although, importance and reliance on $\mathcal{R}_0$ have been debated, it still plays an important role in the public health management, for example, as a measure of infectiousness [28] [38].

In the SIR model defined in system 1.4, using the next-generation matrix approach, $\mathcal{R}_0 = \beta N/\gamma$. Using this formula, $\mathcal{R}_0$ can be interpreted as the product of the expected number of infected at the beginning of outbreak per day βN (the initial force of infection) and the length/period of infectiousness $1/\gamma$. Another way to compute this quantity is to use early data from the outbreak: $\mathcal{R}_0 = e^{gT_g} \approx 1 + gT_g$, where g is the growth rate (of the cumulative number of cases, $\ln(I + R)/t$) and T_g is the mean serial interval (the time between an infector having symptoms onset and infectee having symptoms onset) [29]. Note that these two computations, although describing the same notion $\mathcal{R}_0$, may give different values.

The above-mentioned SIR-model by system (1.4) will be also referred as the "standard system" throughout this thesis. Data that corresponds to the number of currently infected people in a day are referred to as I-data. It is computed as the difference between total infected and recovered individuals up to the given day. Data on the number of currently removed people in a day are referred to as R-data. This is the sum of the total and new recovered individuals. Similarly, I-solution and R-solution will denote respective components in the solution of the standard system. Both data and solution can be sought as a vector of a length equal to the duration of the outbreak and components represent a corresponding value in a given day. Note that using prefixes "$I-$" and "$R-$" is not generally a common way of expression in mathematical epidemiology, but they are considered standard in this thesis.

An example of SIR system (1.4) utility is a model of the Eyam Plague in work [37]. It was proposed that plague had come to Eyam village (England) in 1665. The population went from 350 to 83. There were standout points with this outbreak. First, plague was highly infectious and consequently the whole village was affected. Second, burials were recorded by the rector, so we have outbreak data (although only for R). Third, the same rector isolated the village, hence there was no migration or further spread (fixed N). It was reasonable to apply SIR model (1.4) to this outbreak. It provided a close fit to data and helped to estimate parameters [37]. For example using a root mean square error test, it was shown that $\gamma \approx 3$ (time unit is 1 month). Despite the relative simplicity of SIR system (1.4) (with the census population size), it may provide reasonable results in terms of fitting and parameter estimation; however, there should be certain "ideal" conditions to apply this model.

1.2.3 Effective population size in epidemiology

From the previous discussion on epidemiology in section 1.2.2, it can be seen that population sizes play an important role modeling. Although it is possible to obtain the census population size of a city, this quantity might not be reflective of dynamics due to lockdowns, quarantine and disparate subpopulations, as these situations are supposed to reduce the susceptible population size. Hence, we define *the effective population size* N^*, in the epidemiological sense, as the number of individuals that is needed to get the same SIR dynamics as suggested by real outbreak data. Alternatively, it also can be described as the number of individuals involved in a disease outbreak given data (i.e. $S_0 + I_0$, all people who would or could be infected). Note that there are similarities between the established genetics N_e and our new epidemiological N^*:

1. Both are expected to be less than or equal to the real population size N.
2. Both require observation or data to be estimated.
3. Depending on the definition and the parameters in consideration, there are various ways to compute both.
4. Considering the ratio between effective and census populations may help one to understand underlying features and dynamics.

The reason for introducing this notion is that in small (but still notable) outbreaks, the real population size may provide poor fitting (under a parameter optimization procedure) of simple models to the data. We attempted to fit (least square fitting, the details will be given in the next chapter) the standard SIR model to COVID-19 outbreak data in Tianjin with city's fixed census population size. As seen in figure 1.1, we failed and there was a clear mismatch between data and equations that describe corresponding dynamics. Our procedure was as follows. We needed to have computed γ and $\mathcal{R}_0$ in reasonable ranges (around 0.2 and > 1 respectively, in this case). Thus, the only parameter that we could (significantly) change was β. If β was small, it would make a fitted curve that was almost flat (a zero function) as if there was no outbreak. If β was large, we would obtain unreasonable γ and the fitting would still be inadequate. Even when the achieved optimum parameter values were reasonable, the corresponding fitting was not satisfactory, as in figure 1.1. We conclude that it may not be possible to find optimum parameter values (β and γ) so that an optimizing algorithm would converge to provide a decent fitting.

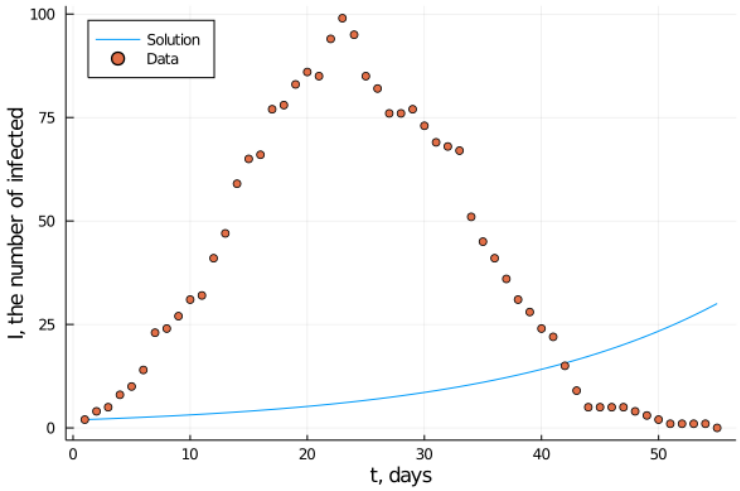


Figure 1.1: Early COVID-19 outbreak in Tianjin with population of 15.6 million. The orange dots represent I-data of the city. The blue line represents the fitted I-component of the standard model with the given population size. $\beta = 2.12 \times 10^{-8}$, $\gamma = 0.28$.

On the other hand, a simple model may not capture all the vital dynamics of the outbreak. At the same time, more complex models where usage of N is feasible may be resource-consuming and may not be easily explainable to the general public. These relatively complicated models for disease dynamics analysis include various methods of extending or transforming a simple SIR model: increasing the number of compartments (e.g. adding Exposed, Quarantine) [3], incorporating features that induce periodicity (e.g. vaccination, seasonality) [14], [2] and introducing heterogeneity in transmission and/or recovery (e.g. age at infection) [43]. We attempt to find whether a simple model with an effective population size can capture outbreak dynamics and relevant specifics. Although more complicated models may provide a better fit, we hope that a simple model with N^* may also give a reasonable fit without adding complexity.

Chapter 2

Computation of the effective population size

In this chapter, we introduce different ways to compute the effective population size, that is, to estimate N^*. In fact, each method may provide an estimate different from others, but this estimate will still satisfy the effective population size definition as in list 1.2.3. In the following section, we derive several methods for computation of N^* that come from a relatively simple model. In section 2.2, we test the derived methods on simulated data in different settings to explore which methods are most accurate in which settings.

Numerical computations are performed using Julia language via VSCode [5] [32]. Statistical analysis is performed using R language via RStudio [36] [40].

2.1 Method discussion and derivation

In this section, we use the standard model introduced in system of equations (1.4) to derive ways of estimating N^*. We treat the total population size N as an unknown constant parameter. We may remove one of the compartments as $N = S + I + R$. Although in most of the mathematical epidemiology literature R is eliminated, as it does not directly play a role in transmission, in our case, we discard S. The reason for such is that we do not know the population size involved initially, so we do not have data for the susceptible class. However, we usually have I-data and R-data. With β and γ having the same meaning as before, we obtain a new reduced system:

$$\begin{aligned} \frac{dI}{dt} &= \beta(N - I - R)I - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{2.1}$$

We have a system of two autonomous differential equations with three parameters. The basic reproduction number is given by $\mathcal{R}_0 = \frac{\beta N}{\gamma}$. We make a substitution of N with N^*, as it is a parameter to be found that does not necessarily match the true population size.

N^* can be found by fitting the system solutions to I-data or R-data using optimization. In Julia programming language [5], we employ the LsqFit package that in turn applies the Levenberg-Marquardt algorithm (LMA) to solve nonlinear least-squares fitting [30] [26]. It is an iterative algorithm that requires an initial point (starting values for parameters to be found) and approaches a local minimum by gradient estimation. At each step, the reduced system (2.1) is solved, one of the solution components is extracted (depending on data to be fit) and LMA attempts to minimize the sum of standard errors (SSE) between the extracted solution and data. During this fitting approach, other parameters β and/or γ could be assumed known or estimated alongside N^*. Alternatively, parameters can be substituted using the $\mathcal{R}_0$ formula, where $\mathcal{R}_0$ is computed using the growth rate, i.e. $\mathcal{R}_0 = 1 + gT_g$.

In addition to fitting methods, there are also transcendental equations for N^* in terms of other parameters such as $\mathcal{R}_0$ and given data-points. In order to state them, we first need to derive an explicit coupled solution of the standard system. We proceed as in [31]. We divide the differential equations for S and I in the system and get a coupled pair of differential equations:

$$\frac{dI}{dS} = \frac{\beta SI - \gamma I}{-\beta SI} \tag{2.2}$$

$$\frac{dI}{dS} = (-1 + \frac{\gamma}{\beta S}). \tag{2.3}$$

Next we integrate and apply the initial condition (S_0, I_0) for an arbitrary constant C:

$$I = -S + \frac{\gamma}{\beta}\ln S + C \tag{2.4}$$

$$I + S - \frac{\gamma}{\beta}\ln S = C \tag{2.5}$$

$$I + S - \frac{\gamma}{\beta}\ln S = I_0 + S_0 - \frac{\gamma}{\beta}\ln S_0. \tag{2.6}$$

Lastly, we substitute $S = N - R - I$ and $S_0 = N - I_0$ into the last equation and obtain the coupled solution in terms of I and R:

$$I + N - R - I - \frac{\gamma}{\beta}\ln(N - R - I) = I_0 + N - I_0 - \frac{\gamma}{\beta}\ln N - I_0 \tag{2.7}$$

$$N - R - \frac{\gamma}{\beta}\ln(N - R - I) = N - \frac{\gamma}{\beta}\ln(N - I_0) \tag{2.8}$$

$$R + \frac{\gamma}{\beta}\ln(N - R - I) = \frac{\gamma}{\beta}\ln(N - I_0). \tag{2.9}$$

Note that we can further substitute $\frac{\gamma}{\beta} = \frac{N}{\mathcal{R}_0}$. If $\mathcal{R}_0$ and a data point ((I, R) value at some t) with non-zero R are given, then it is sufficient to solve transcendental equation (resulting from equation (2.9)) for N^*, as $I_0 \approx 1$ in most cases. Furthermore, in some cases only one of the components is sufficient. If $dI/dt = 0$ (i.e. we have the maximum number

of currently infected individuals, I_{max}), then $S = \gamma/\beta$ and using the (S, I) coupled solution in equation (2.6), we get:

$$I_{max} + \frac{\gamma}{\beta} - \frac{\gamma}{\beta} \ln \frac{\gamma}{\beta} = N - \frac{\gamma}{\beta} \ln (N - I_0). \tag{2.10}$$

Applying the basic reproduction number formula, it reduces to:

$$I_{max} + \frac{N}{\mathcal{R}_0}(1 - \ln \frac{N}{\mathcal{R}_0} + \ln (N - I_0)) - N = 0. \tag{2.11}$$

Note that we do not need information on the number of removed. Equation (2.11) will be referred as the I_{max} formula. On the other hand, a transcendental equation that does not include the I-component is the final size equation that is obtained by dividing the (I, R) coupled solution in equation (2.9) by N:

$$\frac{R}{N} + \frac{1}{\mathcal{R}_0} \ln (N - R - I) = \frac{1}{\mathcal{R}_0} \ln (N - I_0) \tag{2.12}$$

$$\frac{1}{\mathcal{R}_0} \ln \frac{N - I_0}{N - R - I} = \frac{R}{N} \tag{2.13}$$

$$\mathcal{R}_0 = \frac{N}{R} \ln \frac{N - I_0}{N - R - I}. \tag{2.14}$$

Using the point $(I, R) = (0, R_\infty)$ we obtain:

$$\mathcal{R}_0 = \frac{N}{R_\infty} \ln \frac{N - I_0}{N - R_\infty}, \tag{2.15}$$

where R_∞ is the final size of an outbreak, the total number of infected (and eventually removed) individuals throughout its course.

In addition to coupled solutions, standard system (1.4) can be reduced to a single differential equation for R. We divide the differential equation for S by the differential equation for R and integrate:

$$\frac{dS}{dR} = -\frac{\beta}{\gamma} S \tag{2.16}$$

$$\frac{dS}{S} = -\frac{\beta}{\gamma} R \tag{2.17}$$

$$\ln S = -\frac{\beta}{\gamma} R \tag{2.18}$$

$$S = S_0 e^{-\frac{\beta}{\gamma} R} \tag{2.19}$$

$$S = (N - I_0) e^{-\frac{\beta}{\gamma} R}. \tag{2.20}$$

We substitute equation (2.20) into $I = N - R - S$:

$$I = N - R - (N - I_0)e^{-\frac{\beta}{\gamma}R}, \tag{2.21}$$

and use equation (2.21) as a substitution for I in the differential equation for R in system (2.1) to obtain:

$$\frac{dR}{dt} = \gamma(N - R - (N - I_0)e^{-\frac{\beta}{\gamma}R}). \tag{2.22}$$

Note that the solution to this differential equation (2.22) can be used for fitting instead of reduced system (2.1), if we have R-data. In doing so, we avoid the extraction of the component solution.

It would certainly be possible to derive further methods not included in the previous discussion, for example, the ones that utilize other modelling techniques. For example, one could use survival dynamical systems and associated algorithms to find posterior distribution for N [22]. Other parameters (in form of distributions) can be found by fitting to well-known distributions, as it is done in [12], where using clinical observations, incubation period distribution was fitted to Γ distribution. Nevertheless, as it was mentioned, we have focused on the methods that are directly derived from the standard model. We summarize all the methods, which will be used in this thesis, in table 2.1. We are going to test those 7 methods on simulated data and apply them to real data.

Note that we denote data available from the first weeks of the outbreak only as partial data. We require that at least 10 days data is available, so that it is possible to compute the growth rate. For the transcendental equations, in most occasions, $I_0 \approx 1$, it is possible to take this value as 0 for big outbreaks (where we expect N^* to be large so that $I_0/N^* \ll 1$) and use corresponding approximation. Throughout this thesis, the methods will be mostly referred to by their short names.

Method (Short name)	**Data-type**	**Parameters**	**Description**
3 parameters to I (Fitting to I)	Whole or partial data	Fit: N^*, β, γ	Fitting the I-solution of reduced system (2.1) to I-data
3 parameters to R (Fitting to R)	Whole or partial data	Fit: N^*, β, γ	Fitting the R-solution of reduced system (2.1) to R-data
2 parameters to I with β substituted using $\mathcal{R}_0$ formula (β via $\mathcal{R}_0$)	Whole or partial data	Fit: N^*, γ, fixed: $\mathcal{R}_0$	Fitting the I-solution of reduced system (2.1) to I-data with $\beta = \mathcal{R}_0\gamma/N^*$, $\mathcal{R}_0$ computed using the growth rate
2 parameters to I with fixed γ (Fixed γ)	Whole or partial data	Fit: N^*, β, fixed: γ	Fitting the I-solution of reduced system (2.1) to I-data with γ fixed
1 parameter to I with β substituted using $\mathcal{R}_0$ formula and fixed γ (Combination)	Whole or partial data	Fit: N^*, fixed: γ, $\mathcal{R}_0$	Fitting the I-solution of reduced system (2.1) to I-data with $\beta = \mathcal{R}_0\gamma/N^*$, $\mathcal{R}_0$ computed using the growth rate, γ is pre-fixed
Transcendental equation for maximum infected individuals (I_{max})	2 data-points: (I_{max}, R) and $(I_0, 0)$	Fit: N^*, fixed: $\mathcal{R}_0$	Solving the transcendental equation arising from I_{max} formula (2.11), $\mathcal{R}_0$ computed using the growth rate
Transcendental equation for final size (Final size)	2 data-points: $(0, R_\infty)$ and $(I_0, 0)$	Fit: N^*, fixed: $\mathcal{R}_0$	Solving the transcendental equation arising from final size formula (2.15), $\mathcal{R}_0$ computed using the growth rate

Table 2.1: The methods that will be used throughout this thesis for estimation of N^*.

2.2 Verifying via simulations

Our aim in this section is to see if the methods derived in the previous section are accurate and robust. We apply them to simulated data. We investigate parameter values returned

by the methods and compare them to parameter values used for simulations. Although in real outbreaks $N^* < N$, for simulation with homogeneous mixing, we expect $N = N^*$. This is an uncertainty test in the different methods.

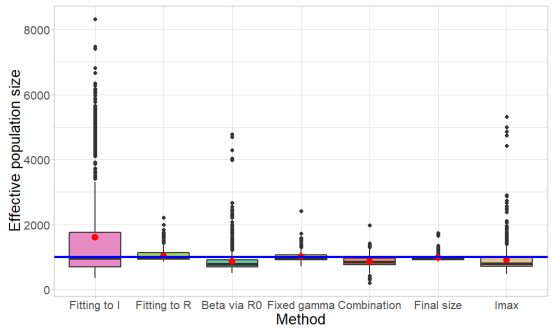
We simulate outbreaks as in [35]. The simulations in this work are based on stochastic Markov chain modeling with exponential infectious periods [17], [48]. We simulate outbreaks by simulating a Markov chain where jumps in a chain correspond to an infection event or a removal event. In this type of process, future outcomes are always based on present ones and those predictions do not differ if instead the whole prior history is used ("memorylessness"). In our case, the standard system (1.4) is a deterministic version of the process, the transitions are the same. A Markov chain consists of tuples of states that are compartments (S, I, R) at time t. Note that although there are three components, there are only two independent variables, since one of them could be represented as a difference of constant population and sum of other two compartment sizes, i.e. $N = S + I + R$.

The time until the next event (a chain jump) is drawn from exponential distribution $T \sim \text{Exp}(\beta SI + \gamma I)$ [48]. The rate parameter of this distribution is composed of infection and recovery rates. There are only two outcomes: a chain jumps to $(S-1, I+1, R)$ with probability $(\beta SI)/(\beta SI+\gamma I)$ (infection) or to $(S, I-1, R+1)$ with probability $(\gamma I)/(\beta SI+\gamma I)$ (removal). Note that transition probability does not depend on time, the process is time homogeneous. The population size is N. An outbreak starts with introduction of 1 infected individual and ends when 0 infected is present. The input is parameter values: β, γ and N, the output is the tuples/vectors of equal length: $\bar{S}$, $\bar{I}$, $\bar{R}$, $\bar{t}$, where each element $S_i/I_i/R_i$ is the number of individuals in the respective compartment at time t_i. Here we present pseudocode for simulation [35]:

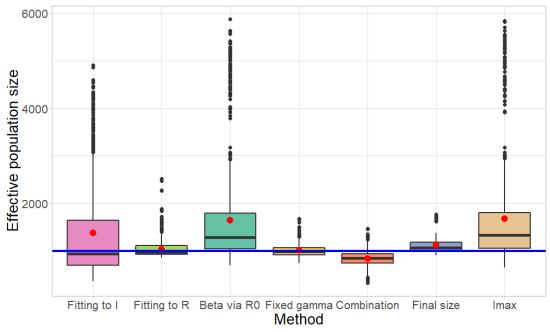
Require: β, γ, N
  $S \leftarrow (N-1)$
  $I \leftarrow 1$
  $R \leftarrow 0$
  $t \leftarrow 0$
  while $I > 0$ **do**
    $t \leftarrow t + Exp(\beta SI + \gamma I)$
    if $Unif(1) \leq (\beta S)/(\beta S + \gamma)$ **then**
      $S \leftarrow (S-1)$
      $I \leftarrow (I+1)$
    else
      $I \leftarrow (I-1)$
      $R \leftarrow (R+1)$
    end if
    Record (S, I, R) and t
  end while

return A list of $\bar{S}$, $\bar{I}$, $\bar{R}$ and $\bar{t}$

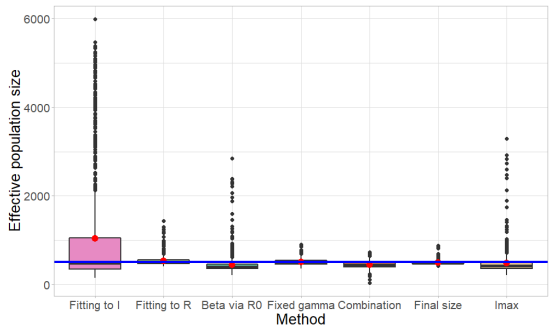
We apply the methods derived in section 2.1 on simulated data and investigate returned parameters. We use several settings varying in N, β, and γ. All the settings have 1000 simulations. The baseline setting is $N = 1000$, $\beta = 0.0005$ and $\gamma = 0.2$, so that $\mathcal{R}_0 = 2.5$. This is close to real outbreaks of a COVID-like disease in terms of the recovery rate and the basic reproduction number [42]. The other settings differ in one or two parameters. Simulations that have a small outbreak ($I_{max} < 50$) or abnormal periods (growth rate=0 for the first 10 days) are dismissed. As for method applications, the starting points for parameters in the numerical optimization are taken to be as true values. The methods that require fixed values such as fixed γ were given true values. For the methods requiring $\mathcal{R}_0$, we used the growth rate over 10 days and serial interval of 6 days (this value is close to the ones of COVID-19) [13]. We focus on the computation of N^*, but we also consider β, γ and SSE between I-data and the corresponding deterministic solution. The results for N^* are summarised in figure 2.1. Results for β, γ and SSE are provided in figures A.1, A.2, A.3 in the appendix.



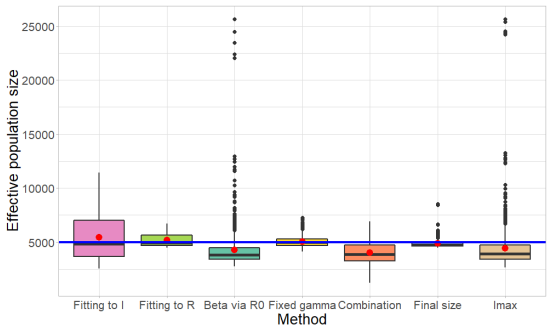
(a) The baseline setting.
$N = 1000$, $\beta = 0.0005$, $\gamma = 0.2$.



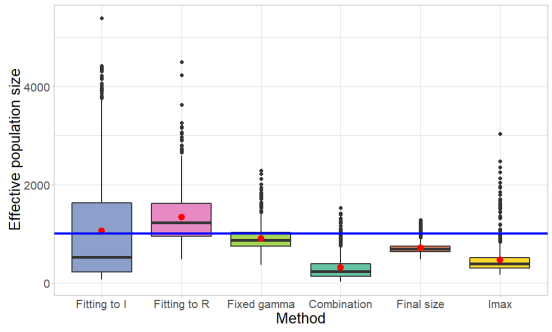
(b) Modified rates.
$N = 1000$, $\beta = 0.00025$, $\gamma = 0.1$.



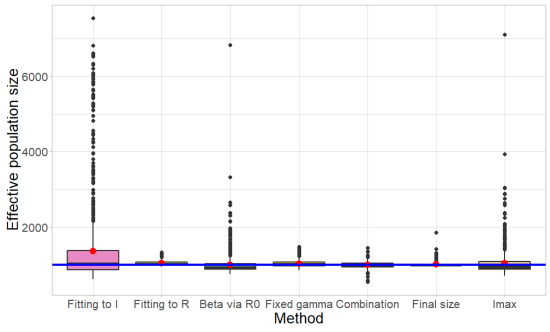
(c) Smaller population size.
$N = 500$, $\beta = 0.001$, $\gamma = 0.2$.



(d) Larger population size.
$N = 5000$, $\beta = 0.0001$, $\gamma = 0.2$.



(e) Smaller $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0003$, $\gamma = 0.2$.



(f) Larger $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0007$, $\gamma = 0.2$.

Figure 2.1: Box-plots for values of N^* obtained from the discussed methods applied to simulated outbreaks data (short names are used). The blue line denotes the true value N. The red dot denotes the mean of 1000 simulations. "Larger" or "smaller" are with respect to the baseline setting parameter.

We use box-plots to analyse how the methods recover parameters from stochastic simulations. We define *outlier* as a value beyond the interval $[q_{0.25} - 1.5 * IQR, q_{0.75} + 1.5 * IQR]$, where IQR is the interquartile range, and $q_{0.25}$ and $q_{0.75}$ are 25 and 75 percentiles respectively. Note that not all methods could be implemented to all settings. In particular, we

were not able to use the β via $\mathcal{R}_0$ method when the basic reproduction number is small ≈ 1.5. In particular, unlike other methods, using true values as the starting point for parameters did not work for the β via $\mathcal{R}_0$ method in this setting. Simulations become more stochastic and less smooth due to the lower transmission rate (i.e. removal and infection happen with similar probabilities at the outbreak start). It is challenging to identify the true underlying maximum (=peak) in simulated data. In the least square fitting process, multiple optima can be encountered. It is possible to find an appropriate starting point by grinding through the parameter space, however, it would take additional resources to work with each simulated data separately. In our analysis, we measure the accuracy of methods by how averages of returned parameter values are close to parameters used for simulation.

Fitting to I produces a significant number of outliers in the computation of N^*. This can be explained by the nature of simulated data. Stochastic simulation data might deviate from the respective deterministic solution (the solution to system (2.1) using true parameter values) and fitting to I produces corresponding deviating optimum parameter values. In addition, this method has the most spread out results with the largest IQR. As the population increases, it gets more precise. In addition, it has the smallest SSE by far.

Fitting to R, fixed γ and final size methods provide relatively accurate results. The latter is surprising since the final size method is a transcendental equation (computationally very fast) and requires only two inputs. However, note that it often fails for smaller $\mathcal{R}_0$. Fitting to R performs well probably because of the shape of data and curve. R-data is typically accumulative with a clear saturation threshold, whereas R-solution is a smooth, increasing function with no fluctuations (there are not multiple local maxima and minima). In addition, fitting to R contains a single differential equation, so it is relatively fast compared to methods that have a system of equations. However, if we use optimum parameters from fitting to R to obtain I-solution, then we get poor fitting to respective I-data, if we compare SSE's between I solution and data. The fixed γ method performs the best in terms of giving the least SSE among methods that return the true population size. However, knowing a parameter value beforehand or during emerging epidemics may not be the case.

β via $\mathcal{R}_0$, combination and I_{max} seem to perform worst as their upper quantile lies below the true value. It seems that the methods that heavily rely on $\mathcal{R}_0$ become worse as this quantity gets smaller. (Although it is true for other methods, but not as drastically). Stochastic simulation data may provide aberrant growth rates, hence the basic reproduction numbers are not reflective of whole outbreak dynamics.

As for the other parameters summarised in figures A.1, A.2, A.3, fitting to R and fixed γ again perform best in returning true values of β. However, for γ it is hard to conclude which one works best by looking at the box-plots, as every method give too many outliers and averages do not always match the corresponding true value. Not surprisingly, methods that are centered around the fitting to I-data have the least SSE which calculated between I-solution and I-data.

Modifying the rates does not change the previous findings. However, note that reducing β leads to more dispersed simulated data. Hence all methods have a larger number of outliers. We further investigate these outliers. In addition to the box-plots, we also considered plotting all of the simulations (both data and solutions) on the same grid. More precisely, we looked at outliers and non-outliers. We have two types of multi-plots. The first one focuses on one method for one setting with an example is given in figure 2.2. It can be seen that outliers of data can be different depending on the considered parameter, although, most non-outliers are concentrated around the corresponding deterministic solution.

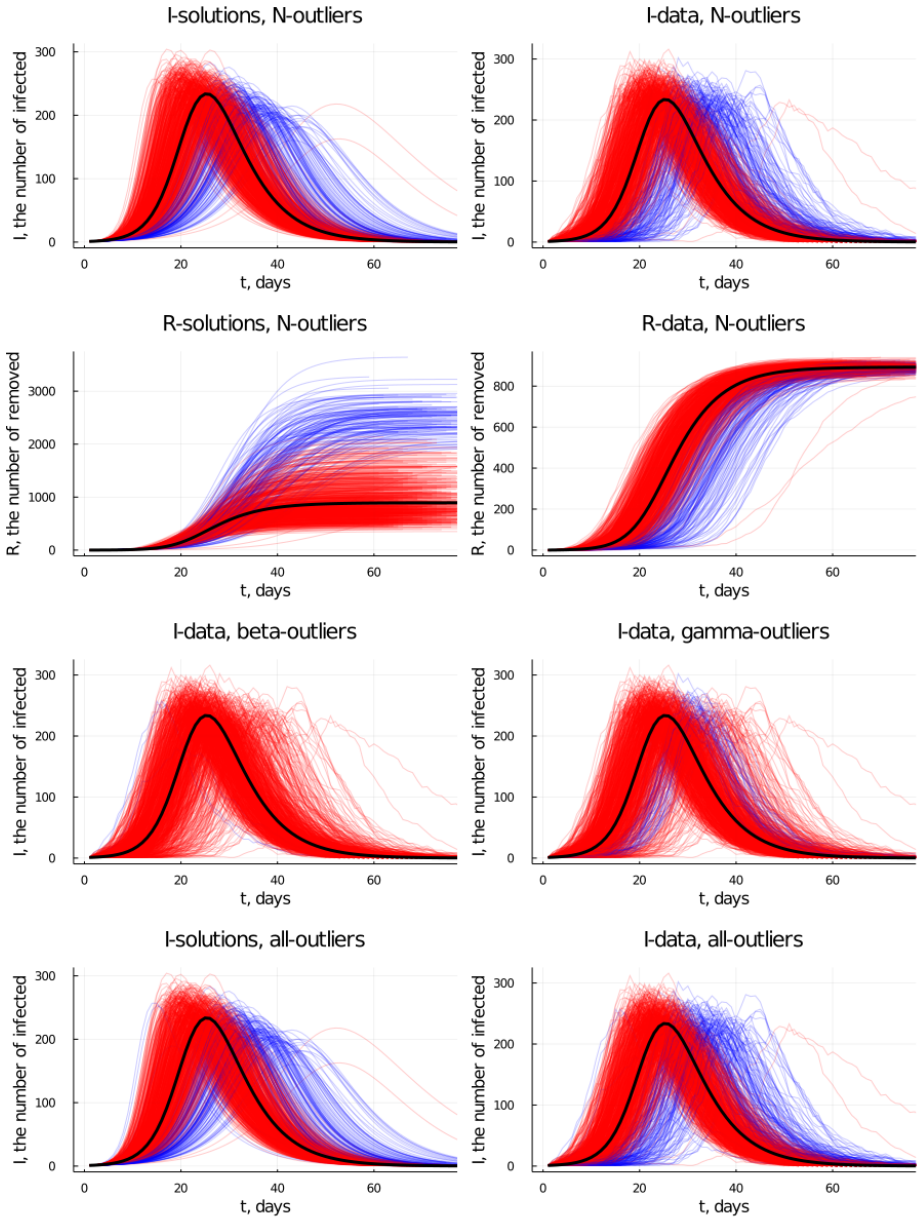


Figure 2.2: Fitting to I-data applied to 1000 simulations in the baseline setting: $N = 1000, \beta = 0.0005, \gamma = 0.2$. Curves are either simulated data '-data' or corresponding deterministic solutions '-solution' that use parameters computed by the method. Blue denotes outliers (data that provided a parameter value beyond the interval $[q_{0.25} - 1.5 * IQR, q_{0.75} + 1.5 * IQR]$ or a corresponding solution), red denotes non-outliers. The deterministic solution with true parameters is shown by the black curve.

For the second type, we look purely on data, but across all methods and settings for one of the parameters. We compared all of the methods, although combination and I_{max} methods have unsatisfactory performance based on the previous box-plot analysis. Those plots are in appendix on figures A.4, A.5, A.6, A.7, A.8. Decreasing β leads to more deviating

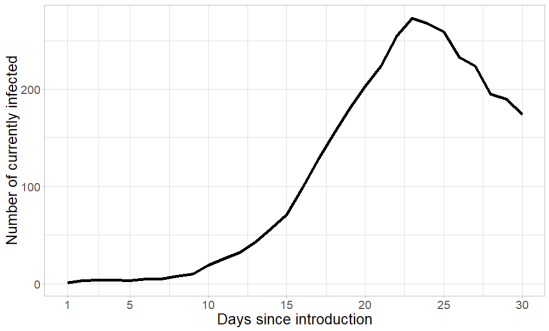
(from the corresponding deterministic solution) data, whereas for increased population size and $\mathcal{R}_0$ all methods have their outliers tightly around the true solution. The former could be explained by the way how the data simulated. β is the most consistent parameter in a way that it is easy to identify the band of non-outliers, it does not have deviating non-outliers. It seems that there is a specific period and timing, outside of which everything is considered to be an outlier. On the other hand, we find γ is the least consistent in terms of having a clear band of non-outliers. For some settings such as a smaller population size, there are two bands of non-outliers: one is around the deterministic solution and the second one is later, shifted by approximately 20 days. N^* seems to combine properties of the above parameters. It is possible to discern the band of non-outliers as in β case, but one could find few deviating non-outliers as in the γ case. The fixed γ method provides the most distinguishable band across parameters and methods: all non-outliers are concentrated around the deterministic solution. The opposite is true for the β via $\mathcal{R}_0$ method: there are a considerable number of deviating non-outliers.

In addition, we look at the performance of the methods in situations when only initial data is available. This is a realistic scenario since, at the start of the outbreak, not all information is available yet, but the demand for estimates of key parameters can be high. We estimate the effective population size using only data from the first days of outbreaks where the true values are the same as in the baseline setting. Another realistic scenario we consider is the case where the contact rate is reduced after some period of time, for example, lockdown implementation. We look at the performance of the methods in this situation. Note that not all methods can be applied. Outbreaks are simulated from the baseline setting. β was halved when the lockdown was implemented. We only look at the results for the population size.

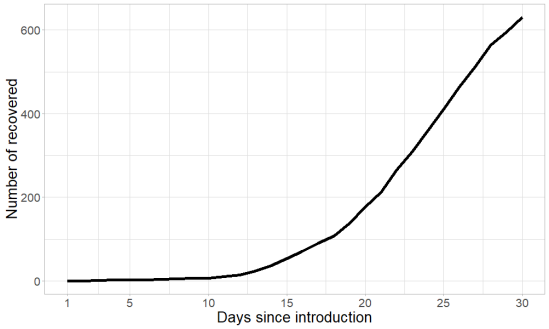
Examples of partial and reduced rate data are given in figure 2.3. There are 1000 simulations. Pseudocode for lockdown strategy is given below.

Require: β, γ, N, *lockdown*
  $S \leftarrow (N-1)$
  $I \leftarrow 1$
  $R \leftarrow 0$
  $t \leftarrow 0$
  while $I > 0$ **do**
    if $t >$ *lockdown* **then**
      $\beta_{used} \leftarrow \beta * 0.5$
    else
      $\beta_{used} \leftarrow \beta$
    end if
    $t \leftarrow t + Exp(\beta_{used} SI + \gamma I)$
    if $Unif(1) \leq (\beta_{used} S)/(\beta_{used} S + \gamma)$ **then**

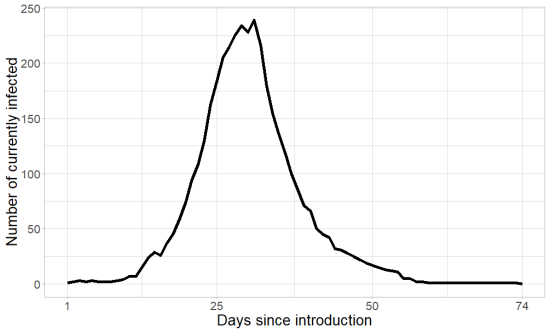
$S \leftarrow (S-1)$
$I \leftarrow (I+1)$
else
$I \leftarrow (I-1)$
$R \leftarrow (R+1)$
end if
Record (S, I, R) and t
end while
return A list of $\bar{S}$, $\bar{I}$, $\bar{R}$ and $\bar{t}$



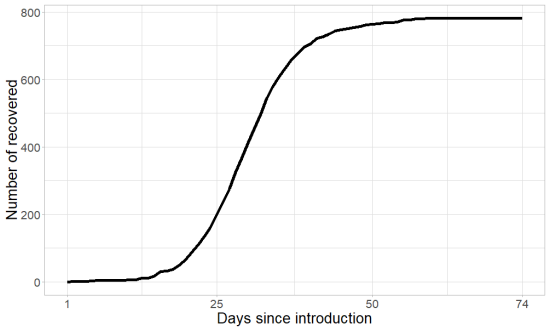
(a) Example of partial I-data of 30 days.



(b) Example of partial R-data for 30 days.



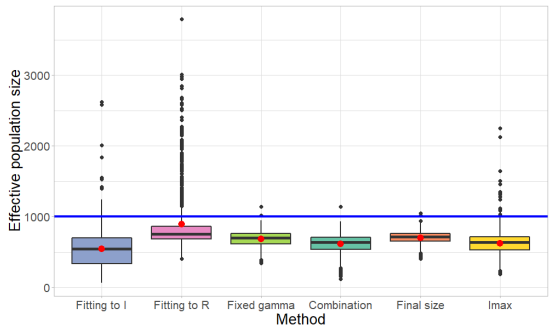
(c) Example of I-data with lockdown implemented after 30 days.



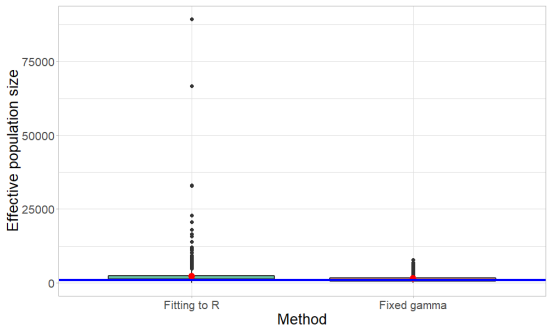
(d) Example of R-data with lockdown implemented after 30 days.

Figure 2.3: Examples of partially available simulated data and simulated data under lockdown. $N = 1000$, $\beta = 0.0005$, $\gamma = 0.2$.

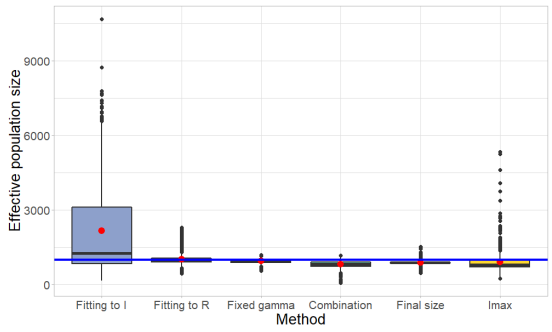
Note that data could be cut after I reached maximum and lockdown could be implemented after cases were plateauing, as in figure 2.3. We aim to show various fixed period lengths and uniformity of what is done to data/simulation. If one wants to consider the exponential growth phase only, then partial data of first 20 days or data of lockdown after 20 days are the most representing in this case. The box-plots of the resulting N^* estimates for both situations can be found in figure 2.4.



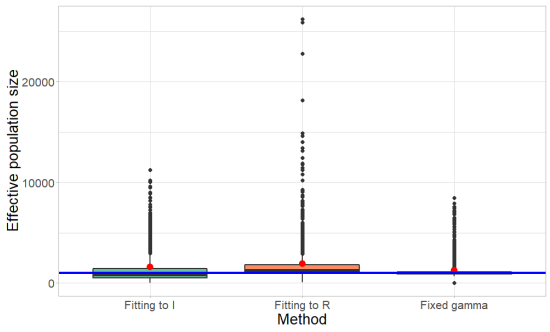
(a) Partial data of first 20 days.



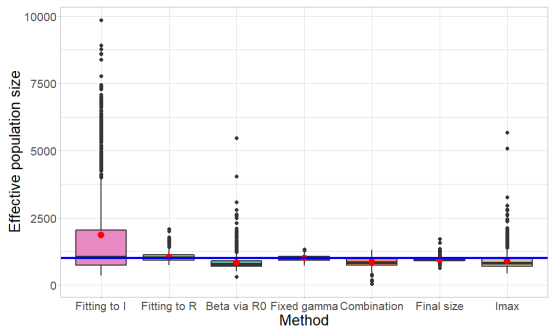
(b) Lockdown after 20 days.



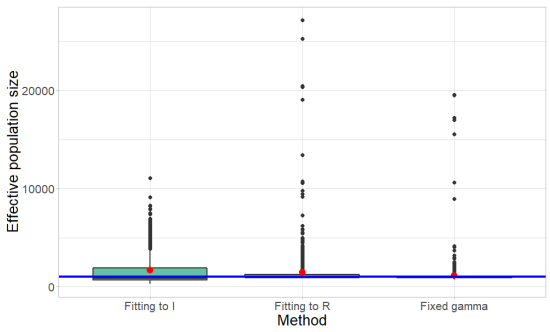
(c) Partial data of first 30 days.



(d) Lockdown after 30 days.



(e) Partial data of first 40 days.



(f) Lockdown after 40 days.

Figure 2.4: Box-plots for values of N^* (y-axis) obtained from the discussed methods under varying levels of partial or lockdown data. The blue line denotes $N = 1000$. The red dot denotes the mean.

If only partial data is available, fitting to R and fixed γ methods still perform the best. However, note that it might be the case that γ could be unknown, so the latter method might not be applicable. In addition, surprisingly, the final size method works as well (in comparison with the others).

If a lockdown strategy is implemented, then we are able to use only a few methods, since we could not apply a single starting point for parameters for a method. In particular, fitting to R and fixed γ can be applied, which further confirms their robustness in different

settings. It is worth noting that fitting to I in some cases (for example, when β is reduced after 30 days) can perform on par with the above-mentioned two.

In conclusion, depending on the known parameters and the goal, there are various ways to perform computations. To obtain the best fit to the infectious data, one needs to employ curve fitting with all parameters unknown i.e. the fitting to I method, as it gives the least SSE between I-data and I-solution. If $\mathcal{R}_0$ is suspected to be relatively high, for example, > 3, one can use a simple final size formula to find N^* or its lower bound. If γ is known, fixed γ can be used to find N^*. Although this method gives higher SSE than fitting to I, fits are still comparable. Otherwise, one needs to use fitting to R to find N^*. Fitting to R and fixed γ methods can be used to find β as well, figure A.1. They are also most robust for partial data and lockdown implementation scenarios, figure 2.4. γ is unpredictable, we could not conclude which method is the best, figure A.2. In fact, none of the methods that derived in this thesis may work to find the removal rate. One has to look at the other ways to estimate this parameter by employing clinical data. For example, one can use data of hospital patients with specifics on when they were admitted and discharged, fit the corresponding time-delay distributions to appropriate probability distributions (e.g. Γ distribution) and take the corresponding mean as the removal rate [12].

Although some methods such as β via $\mathcal{R}_0$ and I_{max} performed inadequately, we are going to apply all the discussed methods to the real data in the next section. Since at the beginning of the outbreaks the basic reproduction number is relatively high, those two methods could still provide some satisfactory results. Real data could be less stochastic in nature, for example, there may be one clearly defined local maximum.

Chapter 3

Application to outbreaks of COVID-19 in Chinese cities

In December 2019, the first known case of pneumonia caused by the novel coronavirus (COVID-19) was identified in Wuhan — one of the major cities of China [49]. By the end of January 2020, cases were reported at other megalopolises and agglomerations such as Chongqing, Beijing and Shanghai. At the same time, the government implemented lockdown and provinces activated public health emergencies [50]. The list of actions included systematic monitoring, travel ban, and making people completely housebound [27]. Despite such measures, COVID-19 spread to other countries and was declared a global pandemic by the World Health Organization in March 2020 [34].

One of the features of this pandemic is the public availability of data on infected, recovered and deceased individuals at various locations. Although populations in Chinese cities can be in several million, the sizes of outbreaks were relatively small. It was seen in figure 1.1 that using the census population size for fitting purposes of the standard SIR model may be inadequate. Thus, we suspect that outbreaks occurred within respective effective population sizes. We apply the methods from the previous chapters to find the effective population sizes in the early outbreaks in selected Chinese cities. In addition, we attempt to find the most favourable model in terms of the fitting to data.

3.1 Data selection

We apply our N^* methodology to data on COVID-19 cases in Chinese cities starting January 15th 2020 till approximately the end of April 2020. This period corresponds to the early outbreaks. The datasets on cities are obtained from Harvard Dataverse [24]: daily "confirmed", "recover", and "death". Each gives corresponding cumulative numbers. The daily number of currently infected people is obtained by subtracting "recover" and "death" from "confirmed". The data is further refined to include only the first outbreak in each city. We concatenate data such that the first day (day 1) corresponds to the first case in the city

and the last day corresponds to when the number of currently infected reached 0. Cities with the following features were excluded:

1. The maximum number of currently (as of January-April 2020) infected was less than 50 ($I_{max} < 50$), since we are looking for outbreaks with a significant number of cases. The outbreaks with small I_{max} are often short, so the growth rate cannot be computed. In small size outbreaks, stochastic processes can dominate transmission, thus our deterministic model methods might not be applicable. We chose to apply methods to datasets, where I-data is similar to a bell-shaped curve and this threshold on I_{max} allowed us to filter such cases. Note that we do not claim that our methods are not applicable in small outbreaks, but we decided to work with real data that is on par with simulated data in section 2.2 for consistency. Almost all of the cities (≈ 270) were rejected because of this criterion.

2. Partial or anomalous data. There are cities that had considerable outbreaks, however, we were not able to complete or amend data. By amend, we mean to make it look like a bell-shaped graph. This would involve adjusting values and/or interpolating, however, there is no obvious way to do this change. These cities are Jining and Wuhan.

3. A first outbreak did not end within our time frame of January-April. A city might have experienced several waves thus the number of currently infected never reached 0 value. These cities are Beijing, Shanghai, Guangzhou, Shenzhen, Foshan and Chengdu.

4. The major outbreak happened after April. For uniformity of conclusions, we excluded cities that had delayed outbreaks. These cities are Dalian and Wulumqi.

5. No information on census population and/or city status. Our approach requires census population sizes N to compare with computed N^* and we were not able to confirm that those locations are indeed city-like. By city-like we mean that a location is densely populated, centralised and has systems for housing, work, transportation, education etc. These locations are Enshitujiazumiaozu, Hubei (Direct Units), Munidiqu, Ganzicangzu.

In total, we have selected 53 cities that did not have any of above issues. The geographic locations can be found in figure 3.1. Plots of outbreak data can be found in figures 3.2, 3.3. Note that the outbreaks started between January 21st and January 26th in the selected Chinese cities, although, it may seem they started at the same time.

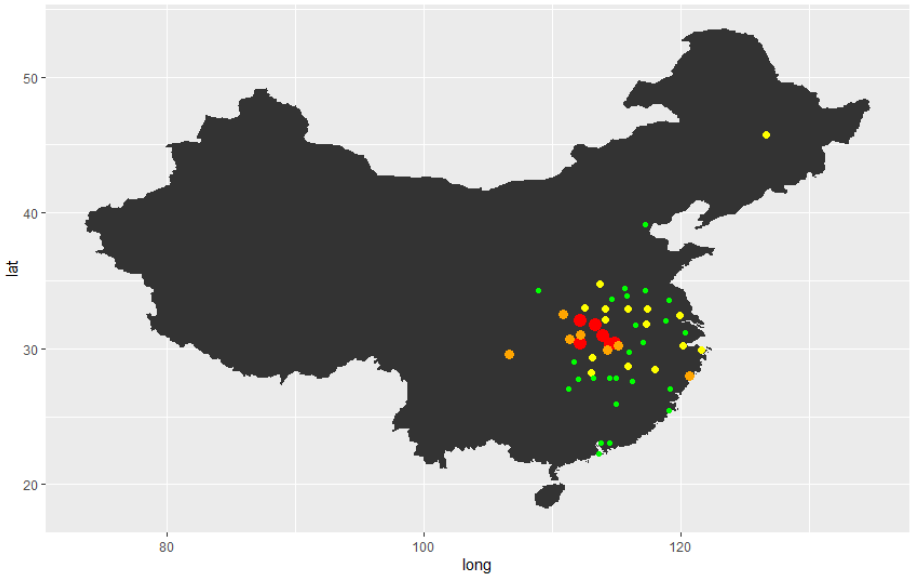
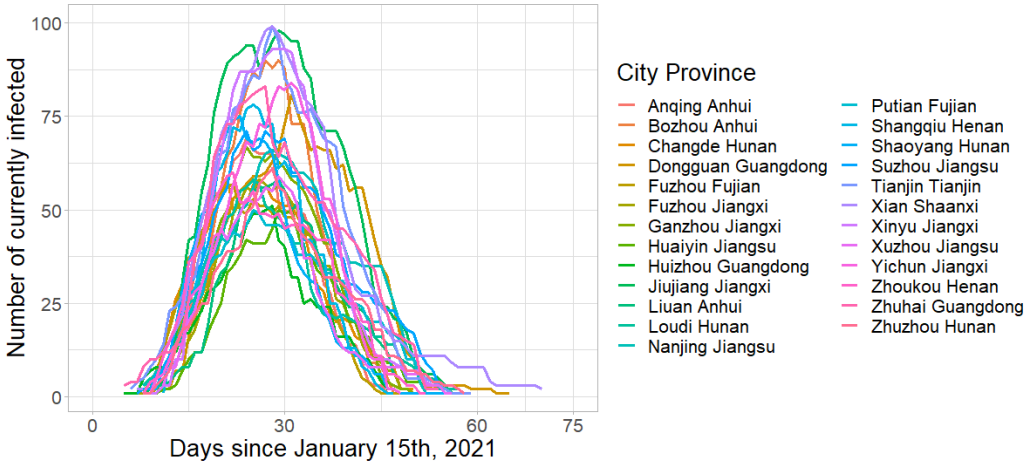
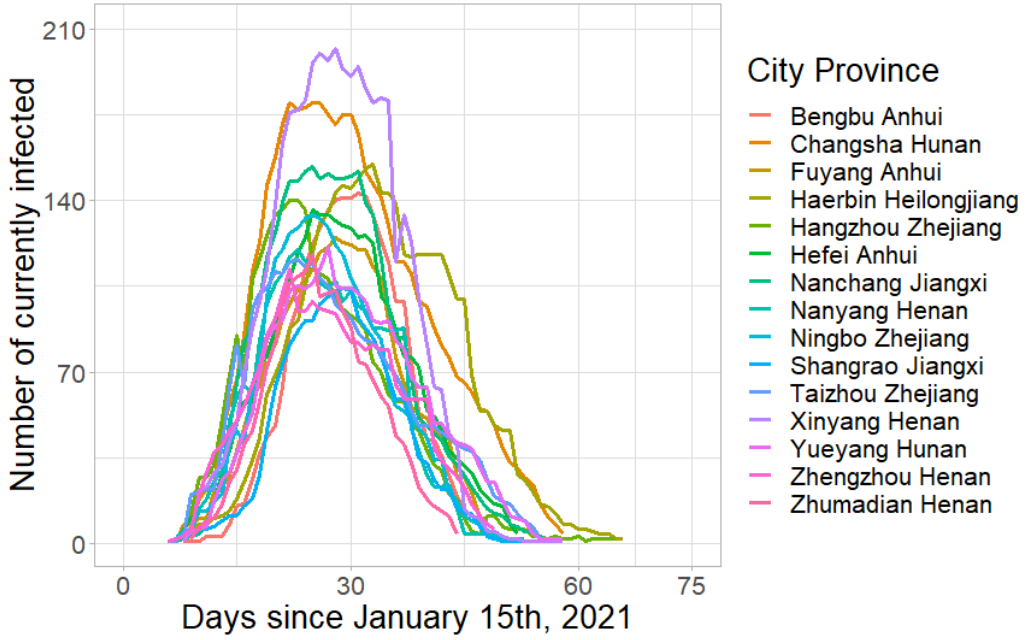


Figure 3.1: Location of 53 selected cities in China. Colors stand for the maximum number of currently infected individuals in a day (I_{max}): green 50-99, yellow 100-249, orange 249-999, red 1000+.

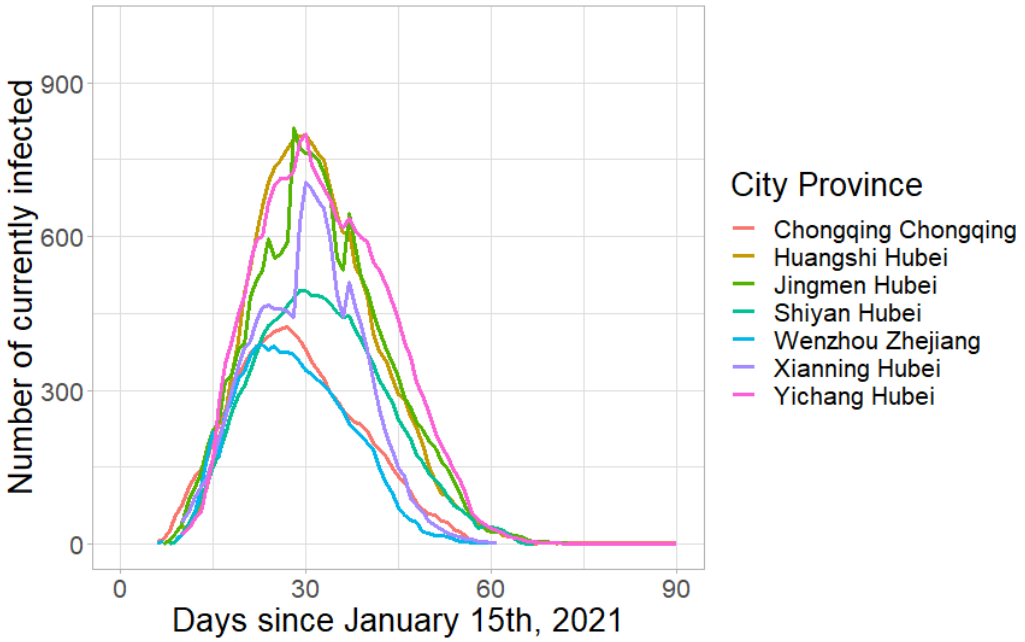


(a) Outbreaks in cities with $I_{max} \in [50, 99]$.

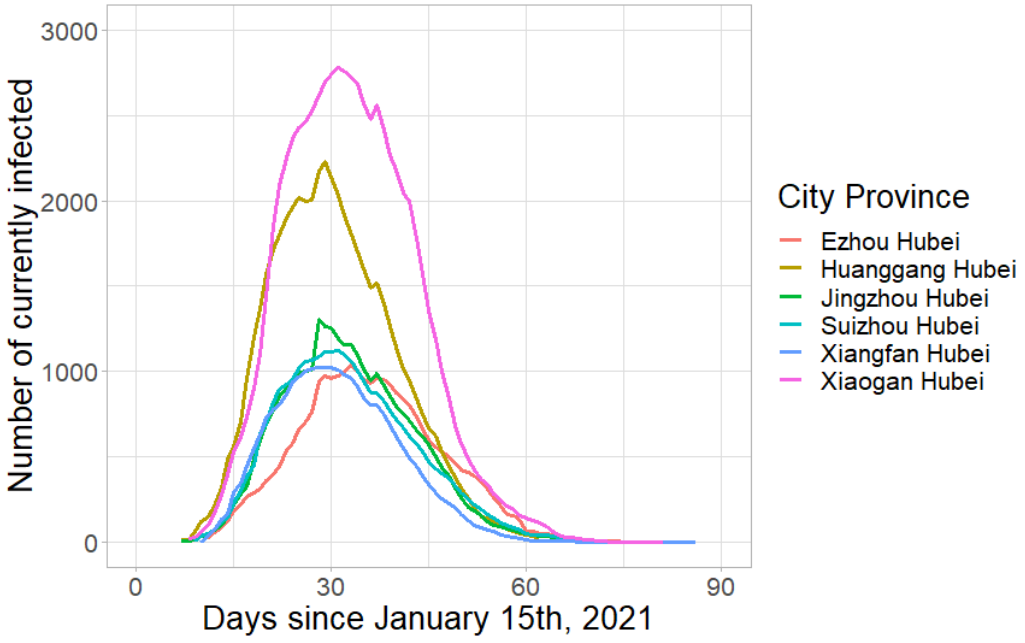


(b) Outbreaks in cities with $I_{max} \in [100, 249]$.

Figure 3.2: I-data for 53 cities divided into groups based on I_{max}.



(a) Outbreaks in cities with $I_{max} \in [250, 999]$.



(b) Outbreaks in cities with $I_{max} \in [1000, \infty)$.

Figure 3.3: I-data for 53 cities divided into groups based on I_{max}.

3.2 Application of methods to estimate the effective population size

We apply the methods derived in chapter 2 to each of the 53 selected cities. For the fixed γ method, we use $\gamma = 0.167$ that corresponds to an infectious period of 6 days as estimated for COVID-19 in [1]. In order to calculate $\mathcal{R}_0$ (for methods that require this estimate), we use the growth rate over the first 10 days of the outbreak with the serial interval of 6 days [13]. A starting point for parameters in model fitting is taken as follows: $N^* = 1000$, $\beta = 0.0005$ and $\gamma = 0.15$. For transcendental equations, an initial point for N^* is to be taken as $R_\infty + 10$. Figures 3.4, 3.5, 3.6, 3.7 show the result of the computations of N^*. In case of considerable confidence interval (CI) range (beyond set limits), a result is omitted. In this paper, only the fitting to R method produced such abnormal results. Note that transcendental equations do not provide CI, hence, they are also not present in the mentioned figures. All the values and 95% confidence intervals for N^* can be found in tables A.1, A.2, A.3 in the appendix. Similar tables for other parameters β (table A.4), γ (table A.5) are also in the appendix.

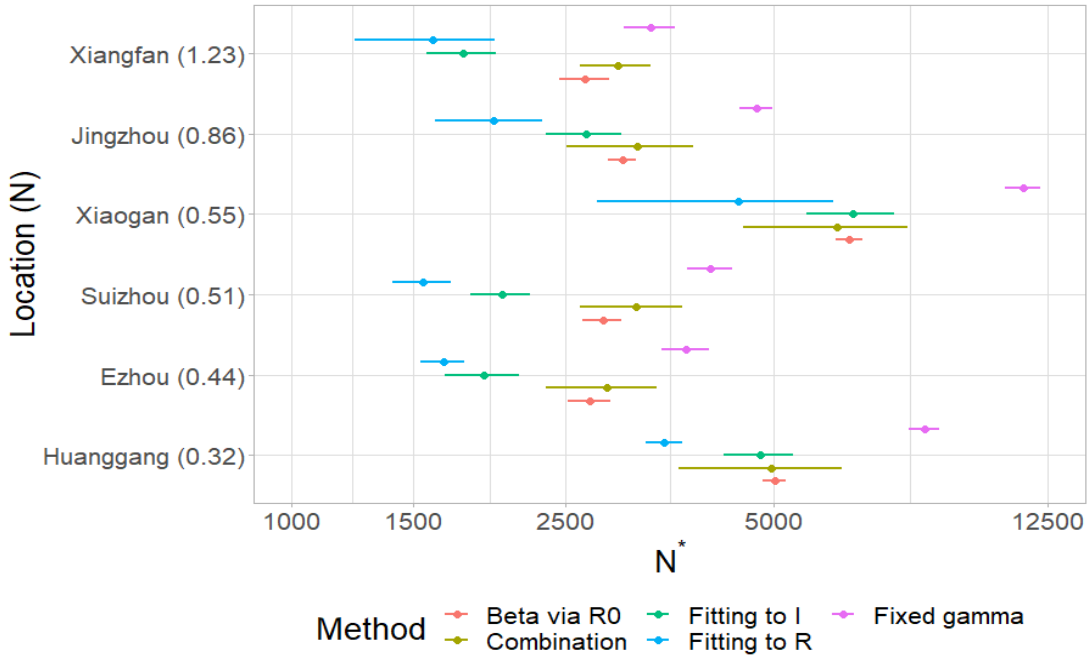


Figure 3.4: N^* results for cities with $I_{max} \in [1000, \infty)$. 95 % CI. x-axis is on a log scale. Numbers in brackets indicate the census population size in millions.

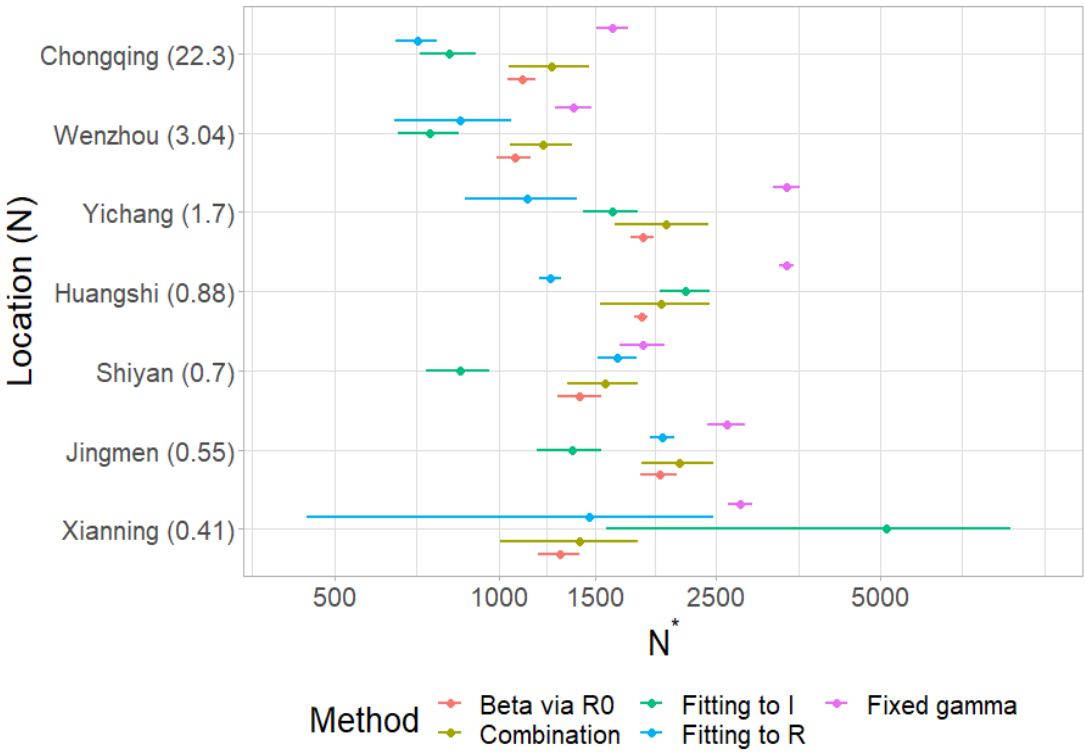


Figure 3.5: N^* results for cities with $I_{max} \in [250, 999]$. 95% CI. x-axis is on a log scale. Numbers in brackets indicate the census population size in millions.

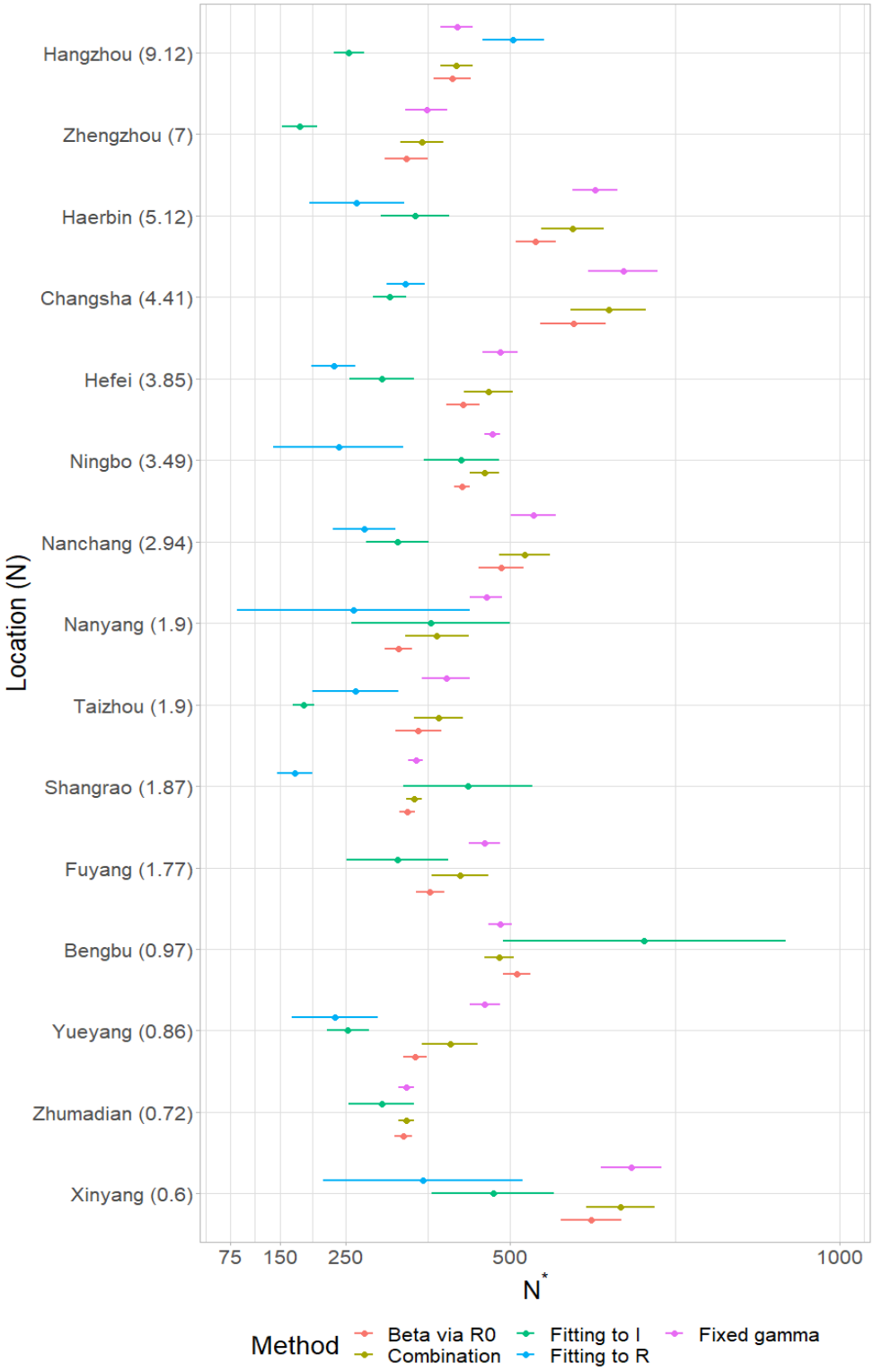
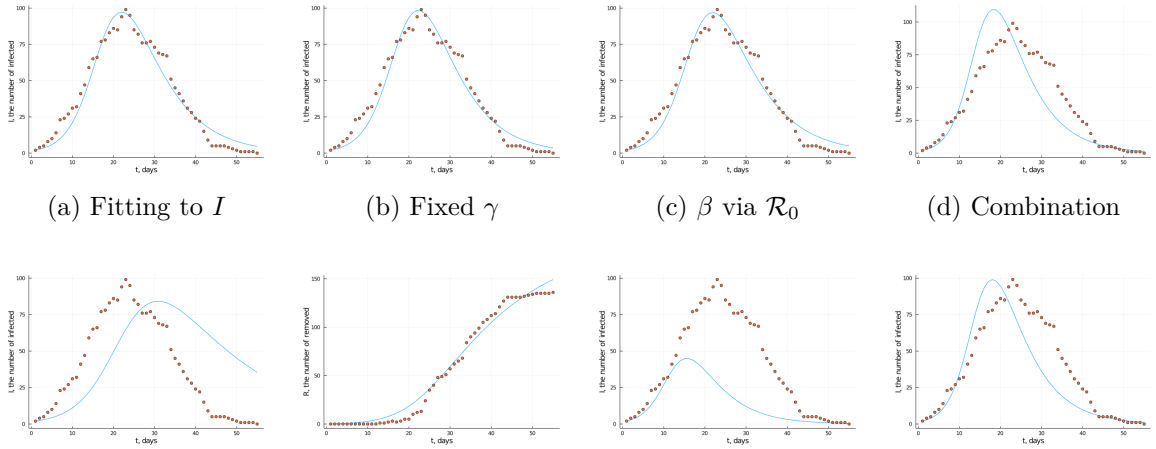


Figure 3.6: N^* results for cities with $I_{max} \in [100, 249]$. 95% CI. Numbers in brackets indicate the census population size in millions. Fitting to R results are omitted for Bengbu, Zhumadian, Zhengzhou, Fuyang.

Figure 3.7: N^* results for cities with $I_{max} \in [50, 99]$. 95% CI. Numbers in brackets indicate the census population size in millions. Fitting to R results are omitted for Liuan, Putian, Shangqiu, Zhuhai, Huizhou, Ganzhou.

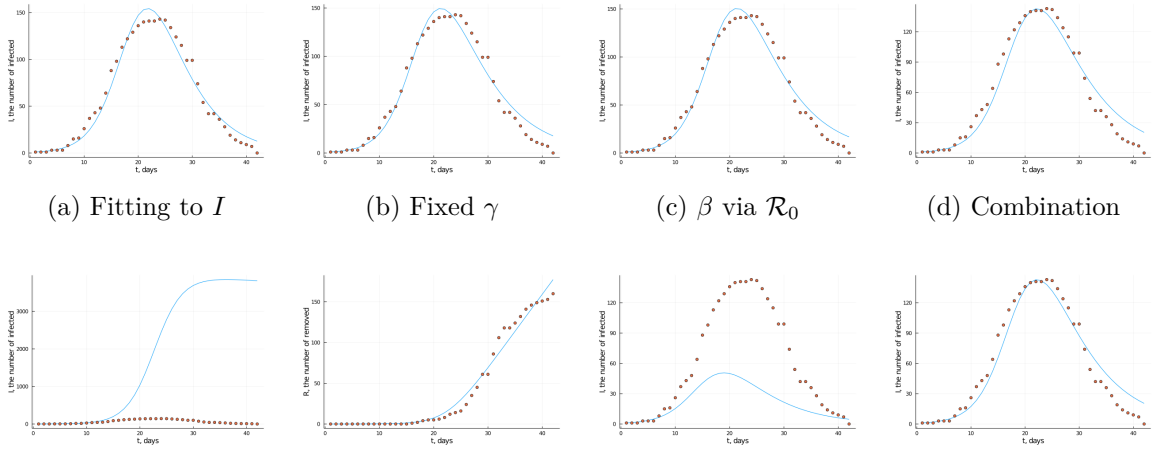
In addition to computation of N^*, we also look at the correspondence between data and solutions. Ideally, we want our method to provide parameters that can be used for curve fitting of the standard model to real data. For each of the methods, we plot I-data and the respective solution for I. For parameters that are not estimated, we use $\gamma = 0.167$ and parametrize $\beta = \mathcal{R}_0\gamma/N^*$, where $\mathcal{R}_0$ is computed using growth rates as defined in chapter 1. Furthermore, for fitting to R, we draw data and a solution for R. Results for three of the cities Tianjin, Bingbo and Yichun can be seen in figures 3.8, 3.9, 3.10. These were chosen to show the variety in how methods that do not include fitting to I-data such as transcendental equations may perform.



(a) Fitting to I (b) Fixed γ (c) β via $\mathcal{R}_0$ (d) Combination

(e) Fitting to R (f) Fitting to R (g) Final size (h) I_{max}

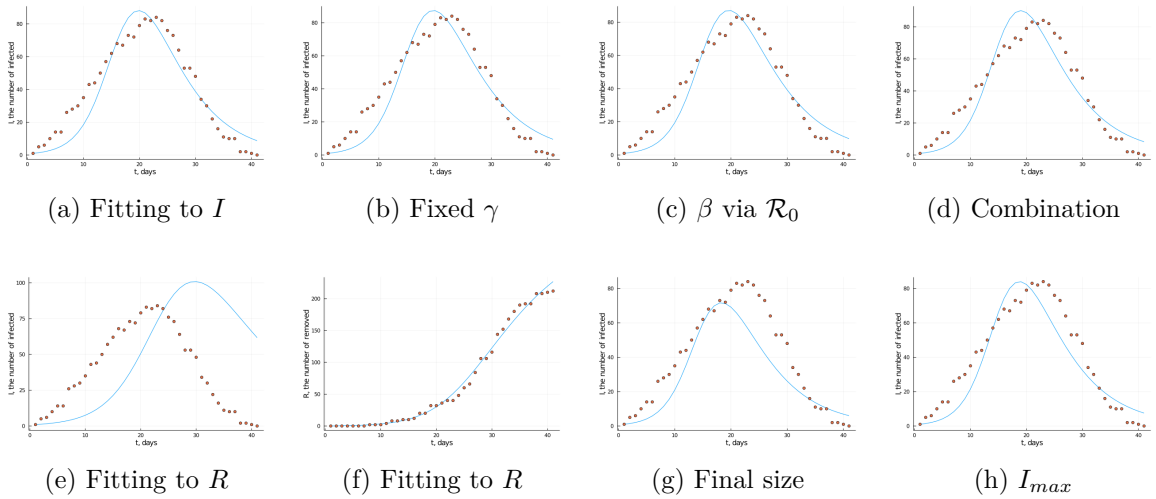
Figure 3.8: Fitting results for Tianjin. Orange dots stand for data. Blue lines stand for corresponding solutions.



(a) Fitting to I (b) Fixed γ (c) β via $\mathcal{R}_0$ (d) Combination

(e) Fitting to R (f) Fitting to R (g) Final size (h) I_{max}

Figure 3.9: Fitting results for Bingbu. Orange dots stand for data. Blue lines stand for corresponding solutions.



(a) Fitting to I (b) Fixed γ (c) β via $\mathcal{R}_0$ (d) Combination

(e) Fitting to R (f) Fitting to R (g) Final size (h) I_{max}

Figure 3.10: Fitting results for Yichun. Orange dots stand for data. Blue lines stand for corresponding solutions.

Based on SSE and plot results, we can see that fitting to I, fixed γ and β via $\mathcal{R}_0$ provide the best fitting. We can further investigate which corresponding model is the best. For this model selection process, we employ Akaike Information Criterion (AIC) [11]. It is an estimator of the relative model quality. Given SSE, numbers of data points and fitted parameters, it provides a score. A model with the lowest score is preferable. In this case, since the number of data points is not large, we will use AICc instead, as it is more sensitive in this type of situation (AIC tends to select models with many parameters in small sample size cases) [8] [16]. The function is then [31]:

$$\text{AICc} = n\Big[\ln\frac{\text{SSE}}{n}\Big] + 2k + \frac{2k(k+1)}{n-k-1}, \tag{3.1}$$

where k is the number of fitted parameters plus one, n is the number of data points. The results are given in the table 3.1. It can be seen that in some cases β via $\mathcal{R}_0$ and/or fixed γ methods have lower scores than fitting to I, although fitting to I offers the lowest AICc.

City name	Fitting to I	β via $\mathcal{R}_0$	Fixed γ
Tianjin	219.99	217.84	219.68
Haerbin	321.94	339.85	349.20
Nanjing	168.95	167.19	174.46
Xuzhou	164.00	161.64	161.75
Suzhou	167.46	213.13	218.85
Huaiyin	117.78	123.76	122.22
Hangzhou	284.67	336.11	337.11
Ningbo	180.95	178.58	181.17
Wenzhou	426.52	447.21	465.29
Taizhou	237.77	308.63	315.19
Hefei	243.09	252.75	259.05
Bengbu	191.02	194.74	196.95
Anqing	164.92	171.66	170.41
Fuyang	215.37	214.06	218.04
Liuan	86.19	84.97	85.70
Bozhou	163.07	160.69	161.15
Fuzhou	189.77	203.09	202.81
Putian	155.97	203.88	203.37
Nanchang	289.29	304.80	310.08
Jiujiang	255.85	283.98	288.96
Xinyu	206.57	207.70	208.51
Ganzhou	164.88	163.45	164.98
Yichun	195.54	193.47	193.27
Fuzhou	112.65	110.67	110.31
Shangrao	173.83	177.41	175.59
Zhengzhou	228.98	260.34	264.05
Nanyang	248.16	246.78	247.17
Shangqiu	128.59	131.76	127.02
Xinyang	341.39	345.14	348.49
Zhoukou	188.49	227.40	227.35
Zhumadian	147.31	146.44	146.90
Huangshi	461.30	476.38	493.85
Shiyan	521.31	551.51	566.83
Yichang	699.51	701.98	750.65
Xiangfan	759.49	792.36	811.11
Ezhou	651.65	675.60	702.32
Jingmen	646.18	663.75	682.65
Xiaogan	833.41	831.21	866.79
Jingzhou	626.19	627.80	662.26
Huanggang	598.65	597.21	640.27
Xianning	440.07	480.68	447.76
Suizhou	594.27	623.23	655.38
Changsha	287.23	363.66	372.18
Zhuzhou	160.61	172.22	174.25
Shaoyang	168.77	173.99	172.05
Yueyang	228.07	244.49	262.08
Changde	169.93	177.35	181.11
Loudi	100.19	105.82	119.47
Zhuhai	175.19	184.01	193.24
Huizhou	107.43	108.05	108.27
Dongguan	268.00	268.63	271.72
Chongqing	406.95	423.64	451.38
Xian	234.93	235.10	244.87

Table 3.1: AICc computed for 3 methods that provide best fitting for chosen 53 cities. Colors denotes the order of scores in each city: highest, middle, lowest.

Furthermore, in order to reassure that our methods indeed provide the lowest SSE between data and the fitted curve, we looked at the respective heatmaps. In each plot, one of the axes is N^* and the other one is the other parameter β or γ. Figures 3.11, 3.12 show one such example. The other cities' and methods' heatmaps are similar and not included. From the first heatmap for example, it can been that the optimization of the SSE is working as intended for the fixed γ method. The respective SSE surface is smooth with no multiple extrema. However, we can not draw the same conclusion for fitting to R method. First, we see a clear band (the dark region in figure 3.12) where multiple parameter tuples can give low SSE. Second, we see that even fixed at 1 value of β, there are other parameter combinations that can provide as low SSE as the optimum parameter do (or maybe even lower). We cannot be sure that we are indeed obtaining the lowest SSE possible when using the fitting to R method, unlike methods that fit a curve to I-data such as the fixed γ one.

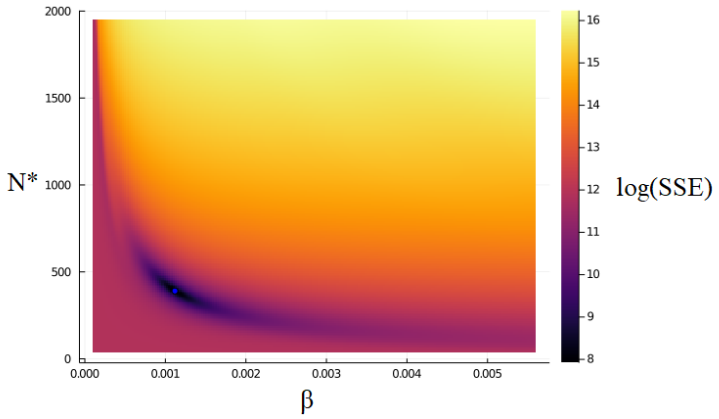


Figure 3.11: A heatmap of SSE for fixed γ method, Tianjin. SSE is between I-data and I-solution. γ is fixed at 0.167. The blue point denotes the parameter values computed by this method, $(\beta_{optimum}, N^*_{optimum})$. Parameter ranges are from 0.1 to 5 with respect to optimum parameter values computed by fixed γ method, i.e. $\beta \in [0.1\beta_{optimum}, 10\beta_{optimum}]$, $N^* \in [0.1N^*_{optimum}, 10N^*_{optimum}]$. SSE is on the log scale. The red contour denotes $log(SSE)$ that is 1.01 multiple of $log(SSE)$ produced by the blue point (not present, since there are no parameter combination giving lower SSE).

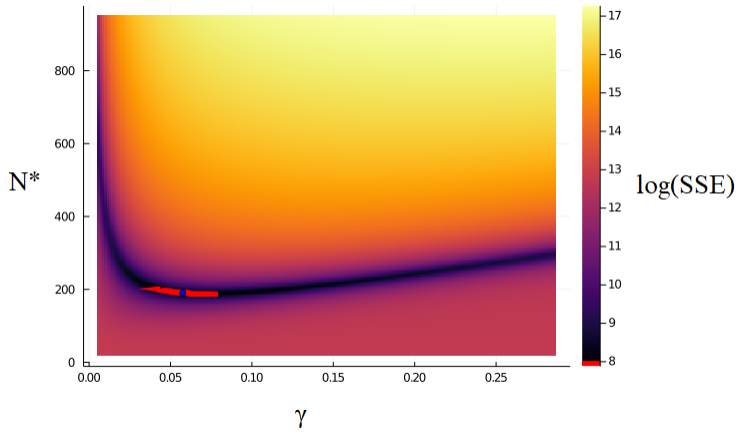


Figure 3.12: A heatmap of SSE for fitting to R method, Tianjin. SSE is between R-data and R-solution. β is fixed at the optimum value computed by this method, $\beta_{optimum}$. The blue point denotes the parameter values computed by this method, $(\gamma_{optimum}, N^*_{optimum})$. Parameter ranges are from 0.1 to 5 with respect to optimum parameter values computed by fixed γ method, i.e. $\gamma \in [0.1\gamma_{optimum}, 10\gamma_{optimum}]$, $N^* \in [0.1N^*_{optimum}, 10N^*_{optimum}]$. SSE is on the log scale. The red contour denotes $log(SSE)$ that is 1.01 multiple of $log(SSE)$ produced by the blue point.

3.3 Discussion on results of methods applied to real data

We observe that all methods computed N^*'s that are considerably lower than corresponding census population sizes. Although $\mathcal{R}_0$ from the growth rate is ≈ 3, and we would expect notable outbreaks, we see that I_{max}s and final sizes are small as compared to N. Such discrepancy between scales of N^* and N can be explained by localization of the outbreaks and implementation of lockdowns. There is no fixed ratio or even a range of N^*/N, but we are going to explore the correlation between this fraction and other factors in this section.

We discuss how the methods computed N^* and other relevant quantities when applied to data on Chinese cities. We also refer to their performance on simulated outbreak data in section 2.2.

The final size method produces the lowest estimates of N^*. This is explained by a relatively high growth rate that results in high $\mathcal{R}_0$ on average (>3). If we look at final size formula (2.15), then the bigger $\mathcal{R}_0$ gets, the closer to the final size N^* is. As the final size method produces underestimated values for N^*, fitting to I data with these N^* and fixed parameters are inadequate. We conclude that initial growth rates do not match the final size, and that is reasonable since the lockdown strategy was implemented.

For fitting to R, we notice that in some cases confidence intervals are large with respect to computed values. We observe that lower CI might even be negative. By analyzing SSE heatmap, we suspect that there are other parameter tuples that can be used for fitting to R-data. Compared to methods that involve fitting to I (with fixed parameters or not), there

are no clear regions on R-data heatmaps to claim the minimum. If we apply the fitting to R method and use optimum computed parameters to get the I-curve, we see that it provides an inadequate fit as well, figure 3.9. In figure A.9, we observe that the fitting to R method is the least correlated with other methods (≈ 0.22), whereas the rest of the methods are strongly correlated among each other (≈ 0.9) β via $\mathcal{R}_0$, Combination and I_{max} give similar values for N^*. From simulated data tests we know that none of these methods performs well in populations of larger size (>5000), figure 2.1. In addition, none of them was applicable for simulated data when partial data was available or lockdown was implemented. Nevertheless, N^* computed by these 3 methods sometimes agree with other methods such as fitting to I and/or fixed γ. Parameters computed by β via $\mathcal{R}_0$, when used for obtaining the I curve, provide good fitting to I-data. Moreover, in some cases, AICc scores of β via $\mathcal{R}_0$ were the lowest among other methods that provide a decent fit. If we look at the computation of γ, this method returns values close to 1/6 that is an actual estimate.

Fitting to I, as expected from simulation results, produces the lowest SSE and consequently the best fitting. However, note that based on figures 3.8, 3.9, 3.10, other methods such as fixed γ and β via $\mathcal{R}_0$ provide a decent fit as well. Moreover, this method's AICc score is not always the lowest. Based on tests on simulated data, we know that this method does not return the actual number of people involved in transmission consistently. If we look at N^* of Chinese cities, there are some values that deviate from what other methods suggest. Similarly to β via $\mathcal{R}_0$, fitting to I often returns γ values close to the actual one. The fixed γ method is one of few methods that return the true population size when tested on simulated data. Applying to real data of Chinese cities, we see that it also provides a decent fit on par with fitting to I and β via $\mathcal{R}_0$ methods. Although the fixed γ method does not often have the lowest AICc score, its weight is approximately the same as the other two in consideration. One possible drawback is that we can observe that in some instances, similar to fitting to I, the fixed γ method gives values for N^* that are not close to what other methods suggest, although there is a strong correlation as in figure A.9. Nevertheless, we use this method for further analysis as it clearly performed better than other methods in all applications we have performed so far.

If we want one overall N^* per city, there are various ways to find it. Recall that one of our goals is to have the same SIR dynamics as suggested by outbreak data. One can choose the fitting to I method, as it provides the lowest SSE on average. However, we have observed that this method did not pass the uncertainty test in chapter 2.2 as good as the other methods. β via $\mathcal{R}_0$ is not applicable in all settings. Among the methods that give a good fitting, fixed γ is the only method that does not have the above-mentioned drawbacks. At the same, we are always sure that at least one of the parameters (γ) is within realistic bounds. In this thesis, fixed γ is chosen as the main method of computing N^*.

We explore the potential of N^* computed with the fixed γ method in the investigation of outbreaks in selected Chinese cities. In particular, we search for implications of N^* or

N^*/N, where N is the census population size of a given city. The results can be seen in figure 3.13. Separate plots with nonlinear fitted curves can be found in figure A.10 in the appendix. We find that the ratio N^*/N is reciprocal to the distance between a location and Wuhan where COVID-19 was first reported. The spatial spread of COVID-19 in China has been studied [18], and our findings further signify the role of the geographic position of a location when considering the outbreak size. There are some points that substantially deviate from the curve and their respective clusters (I_{max} intervals). The corresponding cities deserve investigation why this happened. For example, the rightmost blue point in figure 3.13 is Haerbin. One could research why Haerbin is the only such distant city that has N^*/N higher than expected (the corresponding point is well above the hyperbolic curve in figure A.10c) and yet the outbreak in the city ended (unlike Beijing). One of the explanations could be the position of Haerbin in the high-speed railway system. This city is terminal in the network and connected to other cities in Northern China. It has a large number of travellers at its platforms, hence you observe a bigger outbreak and consequently N^*/N within the cluster. At the same time, it is easier to impose a travel restriction by removing a corresponding edge, hence I would reach 0 instead of some non-zero minima. Another example of a deviating point is Ezhou, a contiguous city with Wuhan, which still has relatively small N^*/N. Unlike other neighboring prefecture-level cities of Wuhan such as Huanggan and Xiaogan, the urban area of Ezhou is almost encompassed by rivers. Similarly to Haerbin, it is easier to impose restrictions by shutting down bridges. There are other points in figures 3.13, A.10 that deserve a separate investigation, and this shows how N^* may have an expansive potential in terms of research even with such simple ideas and limited statistics as the distance between cities.

We attempted to find the correlation between N^*/N and other factors such as density, however, we were not able to identify the right quantity for sub-population in a city where an outbreak happened. Since $N^* < N$, we suspect that outbreaks were centralized in one of the districts of the city. Districts within a city have a varying density, and we do not have information on where outbreaks were centralized and what were densities of these locations. Using the average density of a whole city yielded no strong correlation with N^*/N. With more data and detailed census statistics available, one could further analyse and interpret the differences in N^* and/or N^*/N between cities. For example, the correlation between N^*/N and the length of unrestricted transmission (the time between the first case and lockdown implementation in a city) could shed a light on how a lockdown proceeded. We expect that N^*/N would be proportional to the length (opposite to results in figure 3.13). Any point substantially deviating from the curve (supposedly an exponential one) suggests that in this city lockdown was less or more strict in comparison to other locations in a country. Upon gathering such points, it would be possible to analyse what went right and/or wrong in terms of lockdown management and local policies. This shows how N^* can be beneficial in terms of informing public health teams with directions and decisions.

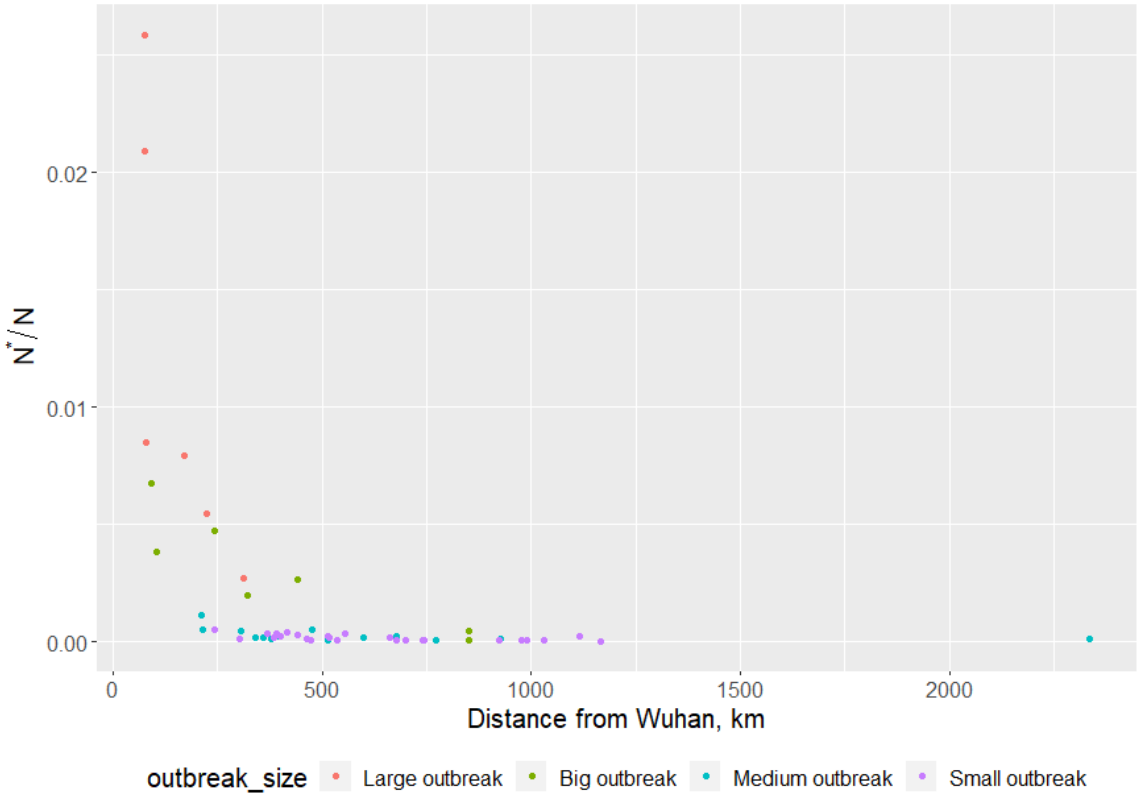


Figure 3.13: The relationship between N^*/N and distance between a city and Wuhan. N^* is computed using fixed γ method (optimum). Colors correspond to the size of outbreak reflected in I_{max}: large $I_{max} \in [1000, \infty)$, big $I_{max} \in [250, 999]$, medium $I_{max} \in [100, 249]$, small $I_{max} \in [50, 99]$.

Chapter 4

Complex model

In this chapter, we extend the standard model in system (1.4) to incorporate heterogeneous population dynamics. The standard model does not explicitly distinguish subpopulations of a given location. Subpopulations can be defined by age, occupation, sub-location etc. Different populations might have different rates of transmission and/or removal. In addition, prevalence, mortality and severity might also vary across subpopulations. An example is childhood diseases such as chickenpox that are prevalent among children but severe among adults. In such a case, one could employ a model that accounts for population heterogeneity, however, we are going to explore if this is necessary. In this thesis, we focus on one such more complex model — a patch model [41]. We stochastically simulate an SIR patch model data and test our methods derived in section 2.1 to estimate the effective population size and relevant parameters. We want to know if the methods that provided reasonable fitting for simple simulation data still work for more complicated ones. We also compare the effective and true population sizes. Since mixing is heterogeneous, we do not expect that all populations in patches (i.e. the sum across the patches) will be involved in disease transmission. In this way, we can explore more strengths and drawbacks of our methods.

4.1 Patch models

A patch model is a metapopulation model, where each subpopulation is a patch and mixing occurs within and between patches [7] [6]. This model can be represented as a graph, where each vertex is a sub-location and each edge is a route between two sub-locations. Subpopulation sizes do not vary across patches (it is not necessarily true in a general patch model), but relevant rates change within and between patches. In our case, each patch may represent a city district and disease dynamics within each patch follows standard model (1.4), although each of the patches, in theory, may follow a different compartmental model. The difference from the standard model is that transmission happens among individuals of the same patch and between individuals from different patches.

$$\begin{aligned}
\frac{dS_i}{dt} &= -S_i \sum_{j=1}^{M} \beta_{j,i} I_j \\
\frac{dI_i}{dt} &= S_i \sum_{j=1}^{M} \beta_{j,i} I_j - \gamma_i I_i \\
\frac{dR_i}{dt} &= \gamma_i I_i.
\end{aligned} \tag{4.1}$$

S_i, I_i and R_i are respective compartments of the i-th patch. There are M of them. γ_i is the removal rate of infected people in i-th patch. $\beta_{j,i}$ is the infection rate between infected of j-th and susceptible of i-th patches. It can be reformulated as $\sum_{k=1}^{M} \beta_k p_{i,k} p_{j,k}$, where β_i is an infection rate of i-th patch, $p_{i,k}$ is a fraction of susceptible from i present in k and $p_{j,k}$ is a fraction of infected from j present k. In this case, it can be viewed as a Lagrangian model [7].

The basic reproduction number $\mathcal{R}_0$ can be computed using the next-generation matrix approach, although the expression gets more complicated as the number of patches increases. In addition, note that each patch has its own basic reproduction number $\mathcal{R}_{0,i}$ that may be different from $\mathcal{R}_0$ of the whole system. $\mathcal{R}_{0,i}$ is approximately $N_i \beta_{i,i}/\gamma_i$ that is the basic reproduction number of the standard model if transmission rates between patches are negligible. This approximation is tighter when transmission between patches gets lower. Throughout this thesis, system 4.1 is referred to as "the complex model".

4.2 Complex model simulation

We simulate from the complex model by extending simulation approach introduced in section 2.2 [35]. Each compartment is a chain. For each patch, $T_{I,i} \sim \text{Exp}(S_i \sum_{j=1}^{M} \beta_{j,i} I_j + \gamma_i I_i)$ is computed, and the smallest non-zero $T_{I,i}$ is chosen as time of a chain jump. Similarly to standard model simulations, there are two outcomes for a patch with the smallest $T_{I,i}$: infection $(S_i - 1, I_i + 1, R)$ with probability $S_i \sum_{j=1}^{M} \beta_{j,i} I_j / (S_i \sum_{j=1}^{M} \beta_{j,i} I_j + \gamma_i I_i)$ and removal with probability $(\gamma_i I_i / S_i \sum_{j=1}^{M} \beta_{j,i} I_j + \gamma_i I_i)$. An outbreak starts with introduction of an infected individual in the first patch ($i = 1$) and ends when none of patches has infected individuals. Pseudocode for the complex model simulation is given below:

Require: $\{\beta_{i,j}\}_{1 \le i \le M, 1 \le j \le M}$, $\{\gamma_i\}_{1 \le i \le M}$, $\{N_i\}_{1 \le i \le M}$
  $(S_1, S_2, \ldots, S_{M-1}, S_M) \leftarrow (N_1 - 1, N_2, \ldots, N_{M-1}, N_M)$
  $(I_1, I_2, \ldots, I_{M-1}, I_M) \leftarrow (1, 0, \ldots, 0, 0)$
  $(R_1, R_2, \ldots, R_{M-1}, R_M) \leftarrow (0, 0, \ldots, 0, 0)$
  $(T_1, T_2, \ldots, T_{M-1}, T_M) \leftarrow (0, 0, \ldots, 0, 0)$
  $t \leftarrow 0$
  while $I_i > 0$ for any $i \in [1, M]$ **do**

for $i \in [1, M]$ **do**

$T_{I,i} \leftarrow Exp(S_i \sum_{j=1}^{M} \beta_{j,i} I_j + \gamma_i I_i)$

end for

Find non-zero minimum $T_{I,i}$ and its index i

$t \leftarrow t + T_i$

if $Unif(1) \leq S_i \sum_{j=1}^{M} \beta_{j,i} I_j / (S_i \sum_{j=1}^{M} \beta_{j,i} I_j + \gamma_i I_i)$ **and** $S_i > 0$ **then**

$S_i \leftarrow (S_i - 1)$

$I_i \leftarrow (I_i + 1)$

else if $I_i > 0$ **then**

$I_i \leftarrow (I_i - 1)$

$R_i \leftarrow (R_i + 1)$

end if

Record $(S_1, S_2, \ldots, S_{M-1}, S_M)$, $(I_1, I_2, \ldots, I_{M-1}, I_M)$, $(R_1, R_2, \ldots, R_{M-1}, R_M)$ and t

end while

return A list of $(\bar{S_1}, \bar{S_2}, \ldots, \bar{S_{M-1}}, \bar{S_M})$, $(\bar{I_1}, \bar{I_2}, \ldots, \bar{I_{M-1}}, \bar{I_M})$, $(\bar{R_1}, \bar{R_2}, \ldots, \bar{R_{M-1}}, \bar{R_M})$ and $\bar{t}$

We consider 11 settings for the complex model simulation. All patches have a population of 1000 and a removal rate of 0.2. We change only infection rates within/between patches. All the main settings are on 3 patches. The transmission matrices and short names are given in table 4.1. Note that some of the matrices are 180 rotations of another, however, the index case is in the first patch, so these cases are indeed different. We have two asymmetric matrices (3.6 and 3.7 in table 4.1) representing cases where transmission rates vary depending on a patch pair, i.e. $\beta_{i,j} \neq \beta_{j,i}$. In addition, we have simulated an outbreak on 5 and 9 patches. Each has only 1 setting and it is an extension of 3.1 in table 4.1. For outbreaks on 5 patches, the transmission rate within a patch is 3 and the transmission rate between patches is 0.5. For outbreaks on 9 patches, the transmission rate within a patch is 3 as well, but the transmission rate between patches is 0.25. These settings are labelled as 5.1 and 9.1 respectively. Since the number of settings grows exponentially with the number of patches, we have not considered other cases, for example, settings involving isolates and clusters. This is a topic for another research, and our aim is to show one of the possible extensions. We use 1000 simulations for each setting and abnormal simulations were not included (see section 2.2 for criteria).

$\begin{bmatrix} 4.5 & 0.25 & 0.25 \\ 0.25 & 4.5 & 0.25 \\ 0.25 & 0.25 & 4.5 \end{bmatrix}$ Basic (3.1)	$\begin{bmatrix} 4.5 & 0.5 & 0.01 \\ 0.5 & 4.5 & 0.01 \\ 0.01 & 0.01 & 5 \end{bmatrix}$ Cluster-in (3.2)	$\begin{bmatrix} 5 & 0.01 & 0.01 \\ 0.01 & 4.5 & 0.5 \\ 0.01 & 0.5 & 4.5 \end{bmatrix}$ Cluster-out (3.3)
$\begin{bmatrix} 4.5 & 0.25 & 0.25 \\ 0.25 & 5 & 0.01 \\ 0.25 & 0.01 & 5 \end{bmatrix}$ Bridge-in (3.4)	$\begin{bmatrix} 5 & 0.25 & 0.01 \\ 0.25 & 4.5 & 0.25 \\ 0.01 & 0.25 & 5 \end{bmatrix}$ Bridge-out (3.5)	$\begin{bmatrix} 6 & 0.25 & 0.1 \\ 0.5 & 4 & 0.1 \\ 0.5 & 0.25 & 2 \end{bmatrix}$ Descending (3.6)
$\begin{bmatrix} 2 & 0.25 & 0.5 \\ 0.1 & 4 & 0.5 \\ 0.1 & 0.25 & 6 \end{bmatrix}$ Ascending (3.7)	$\begin{bmatrix} 7 & 0.01 & 0.01 \\ 0.01 & 5 & 0.01 \\ 0.01 & 0.01 & 3 \end{bmatrix}$ Decreasing (3.8)	$\begin{bmatrix} 3 & 0.01 & 0.01 \\ 0.01 & 5 & 0.01 \\ 0.01 & 0.01 & 7 \end{bmatrix}$ Increasing (3.9)

Table 4.1: Transmission rate matrices $\{\beta_{j,i}\}$ for complex model simulation on 3 patches. Short names and number labels are provided under each. All the values get multiplied by 0.0001.

We apply the methodology derived in section 2.1 (all 7 methods) to cumulative simulated data (the daily sum of cases across all patches). We want to know if the derived methods would give optimum parameter values that in turn provide a decent fit to data. For outbreaks on 3 patches, the standard starting point for fitting is $[\beta, \gamma, N^*] = [0.0005, 0.2, 1000]$. For setting 5.1 and 9.1, the starting points are $[\beta, \gamma, N^*] = [0.0001, 0.2, 5000]$ and $[\beta, \gamma, N^*] = [0.000056, 0.2, 9000]$ respectively. We have a few exceptions. For setting 3.7 we could not use the standard point, instead, we are able to use $[0.0005/3, 0.2, 3000]$. For setting 3.6 and the β via $\mathcal{R}_0$ method we use initial $N^* = 1500$. We are not able to perform certain methods in some settings, because there is no starting point that would work for all 1000 simulations. One has to grind through a parameter space for each simulation separately. In particular, in settings 3.7 and 3.9, we do not have results for fitting to I and β via $\mathcal{R}_0$ methods and similarly in setting 3.8, we do not have results for fitting to R method. Similarly to tests in section 3.2, we obtain optimum values for parameters and SSE, but we choose to focus on N^* for discussion in the next section.

4.3 Complex model results

We discuss computed values of N^* and the fitting potential of derived methods when applied to complex model simulations. Once optimum parameter values are computed, we use box-plots for N^* to compare results across methods and with a true value of N (the sum of 3/5/9 patches). There are thresholds for population sizes on the plots, so not all outliers are present. In addition, we pick one of the simulated datasets and plot the fitted curves

alongside the simulated outbreak. The results and examples are demonstrated in figures 4.1, 4.2 and 4.3. We have limited the box-plot analysis to values within $[0, 2N]$ to look closely at the performance around the true value. Thus, some of the methods appear concatenated. We note similarities with simple simulation results, however, the situation changes once we have more isolated patches and asymmetric transmission rates.

Setting 3.1 is the closest to homogeneous mixing because it is symmetric across patches and transmission rates between patches are comparable to those within. There are no multiple extrema in I-data. Setting 3.4 is almost like homogeneous mixing since the introduction happens in the patch that is most connected to the rest. In both settings, outbreaks happen almost simultaneously in all patches. Hence, we observe that on average N^* is close to N, similarly to simple model simulations. Setting 3.5 is also close to homogeneous mixing as patches are well-connected. There is a short delay in an outbreak in the third patch and hence the one maximum is spread out.

Unlike settings 3.1, 3.4 and 3.5, we see that some methods provide N^* that is clearly less than N. Settings 3.2 and 3.3 have similar results. First, we know that in at least one of the patches (in this case, patch 3) an outbreak may or may not happen (e.g. infection was contained in a cluster) and if it happens it may be delayed, since the transmission rate between patch 1 and patch 3 is low. Hence cumulative I-data have 2 maxima that are still close (settings 3.2, 3.3). In the fitting process, we are not able to capture these nuances. Nevertheless, similarly to setting 3.5, 3 methods (fitting to R, fixed γ and final size) are returning N^* that is close to true size N. In fact, on box-plots, the disparity between methods in providing optimum N^* that is close to N is more obvious than in simple model simulation results, although we do not expect that $N^* \approx N$.

In settings 3.6 and 3.7, although we may observe a typical I-curve with 1 maximum, the underlying dynamics are not that simple. In setting 3.6, we observe that an outbreak passes quickly in the first patch and there is not enough transmission for a full outbreak in the last patch. Hence, a smaller number of individuals are affected. This is the first setting where we can clearly see that $N^* < N$ across all the methods. In addition, this setting is the last one where we observe a discrepancy between methods in terms of how computed N^* is close to N. In setting 3.7, we are not able to use a single starting point for all simulations when applying fitting to I and β via $\mathcal{R}_0$ methods. This can be explained by the fact that outbreaks happen with various delays across simulations. We do not expect an outbreak to progress quick enough in patch 1 on its own since rates are low. Nevertheless, once there is an introduction in patch 3, cases grow across all patches almost simultaneously. This is why we expect to see a simple cumulative I-curve with slow growth and shifted maximum. There are simulations, where a maximum is reached after 10 weeks from the introduction. We observe that all methods provide optimum N^* that is close to N.

When we move to settings 3.8 and 3.9, where patches are more isolated, we observe multiple extrema in I-data. Unlike settings 3.2 and 3.3, here 2 or more maxima can be

distinguishable and far away from each other. Moreover, outbreaks may not happen in all patches (similarly to setting 3.6). In setting 3.8, it leads to the case where only one of the "bumps" is captured when we attempt to fit to data. Consequently, we obtain $N^* < N$. Note that we are not able to apply the fitting to R method in this setting, again due to the absence of a single starting point that would work for all simulations. It may be explained by a large variation in outbreak length. Setting 3.9 is different from setting 3.8, because introduction happens in a patch with a low transmission rate, thus we expect slow growth in the number of cases. Setting 3.9 is similar to setting 3.7 in the way that a global maximum may appear with a considerable delay. Hence, we are not able to apply fitting to I and β via $\mathcal{R}_0$ methods. Even if we can apply a method, none provides a reasonable fitting to data or computation of N^* with exception of the final size method. These two settings are examples where our methodology is inadequate.

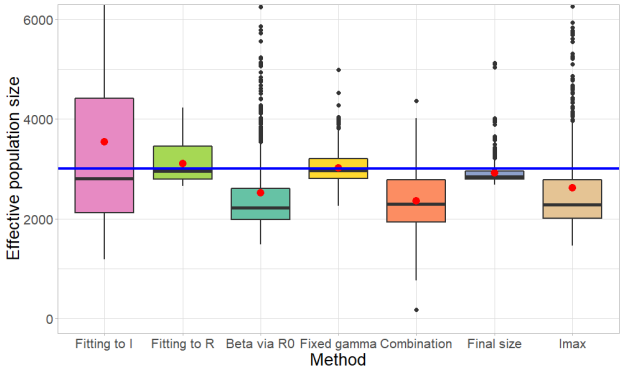
From the previous two chapters, we know that the fixed γ method is the optimal one since it provides a reasonable fit to both simulated and real data and in the case of homogeneous mixing simulations it returns $N^* \approx N$. We observe that in settings 3.1 and 3.4 of complex simulation where one expects a single maximum, this method, fitting to I and β via $\mathcal{R}_0$ still can be used for fitting purposes. In settings 3.2, 3.3, 3.5-3.7, these 3 methods can still provide a decent fit by capturing the largest extrema in I-data. However, if an outbreak has multiple distinct extrema as in settings 3.8 and 3.9, then neither of the methods give optimum parameters for an adequate fitting. Fitted curves either capture only one of the bumps or go through the mean of those bumps. Other methods (fitting to R, combination and transcendental equations) cannot be used for fitting purposes in almost all settings.

When we look at N^* results, we have various conclusions for each of the settings. In settings 3.1 and 3.4 that are close to homogeneous mixing, we get $N^* \approx N$ similarly to simple simulation results in section 3.2. On the other hand, in settings 3.6, 3.8, we clearly see $N^* < N$. The common feature between these 2 settings is that transmission rates decrease from patch to patch. There might not be a fully realized outbreak in the last patch. We conclude that not all individuals are participating in disease dynamics. In settings 3.2, 3.3 and 3.5, there are some methods that provide $N^* \approx N$. It is a surprising result, since, in settings 3.2 and 3.3, mixing is far from homogeneous, whereas in setting 3.5, mixing is close to homogeneous, but not all methods return N. In settings 3.7 and 3.9, where transmission rates increase, we see that N^* may be larger N. This can be explained by the fact that since transmission initially is slow (β is estimated on the lower end), but there is a significant peak in I at some point, methods compensate with higher N^*. To sum up, we observe, except for some methods in a couple of settings, similar results as in the previous sections, N^* is less than N.

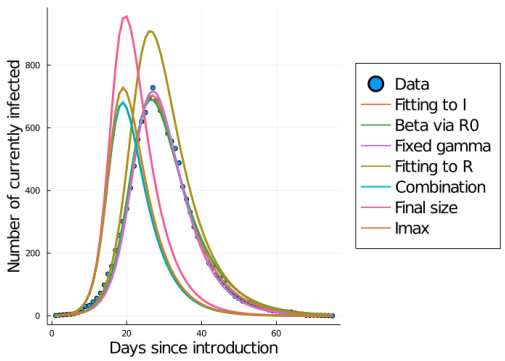
If we look at settings with a larger number of patches, 5.1 and 9.1, we observe that all the conclusions that are drawn for setting 3.1 are true for these settings as well. It is expected since 5.1 and 9.1 are extensions of 3.1. The fixed γ method is still relatively the

best. However, we note that the fitting to I method gets less spread out results. In general, we observe that averages (mean and median) are closer to N as the number of patches and consequently N increase. We have already seen in simple model simulations that the larger N is, the more accurate results are. Our extensions further confirm this conclusion.

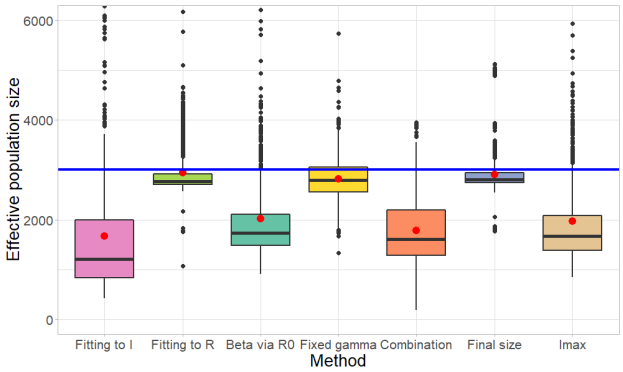
Because we do not (and cannot) know what "true" N^* should be, we may not really assess the performance of the methods in terms of returning optimum N^* as we move to heterogeneous mixing situations. If we still expect N^* to be close to N, then there are methods (including fixed γ) that can accurately capture $N^* \approx N$ in certain settings (3.1-3.5 and maybe 3.7). In general, we have found that our methods that derived from a simple model in system (1.4) are still applicable in some heterogeneous mixing settings. Those settings are the ones where patches are well connected (transmissions rates are not negligible), so there no isolates or isolated clusters. In settings, where none of the methods was able to capture the dynamics, we propose possible ways of dealing with them. One may employ or find another infectious disease model as discussed in chapter 2. Another solution is to dissect data at local minima and reapply our methods to new data. Although finding the optimum way of handling settings where our methods failed is beyond this thesis's goals, this further shows how N^* and corresponding fitting could be a fruitful source for research.



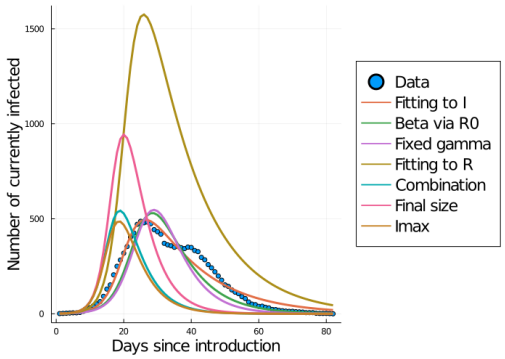
(a) Basic (3.1).
Box-plot results for N^*.



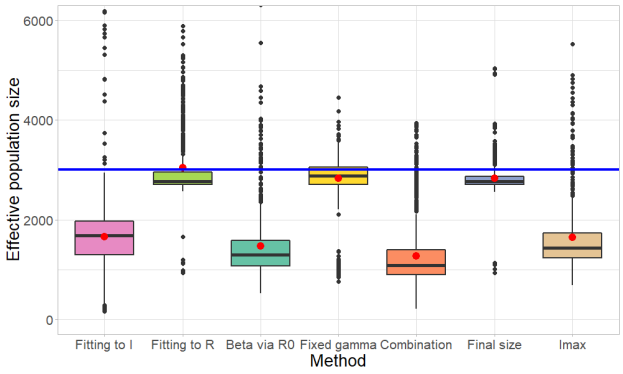
(b) Basic (3.1).
Example of fitting results.



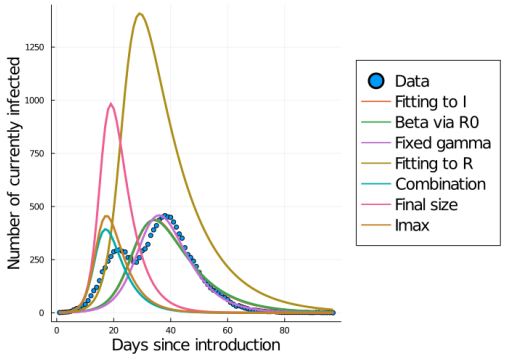
(c) Cluster-in (3.2).
Box-plot results for N^*.



(d) Cluster-in (3.2).
Example of fitting results.

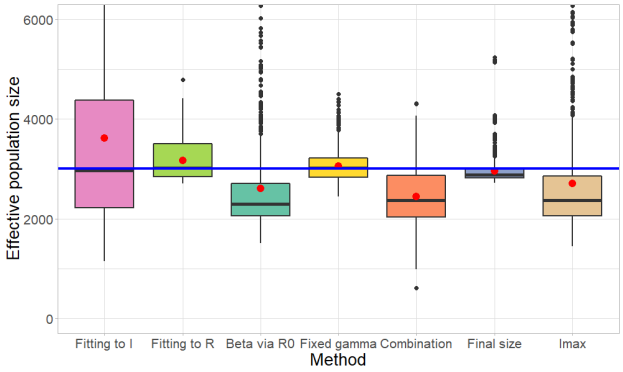


(e) Cluster-out (3.3).
Box-plot results for N^*.

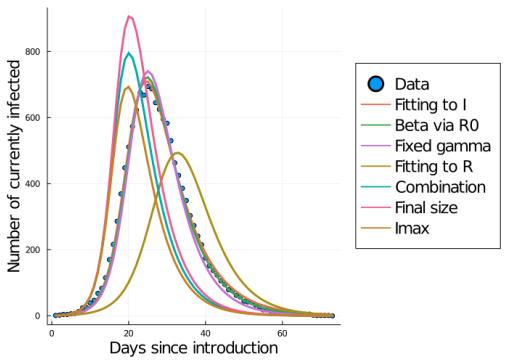


(f) Cluster-out (3.3).
Example of fitting results.

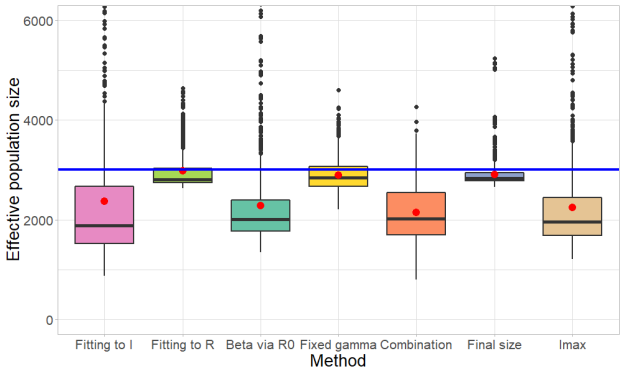
Figure 4.1: On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axis is restricted to $[0, 6000]$. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula.



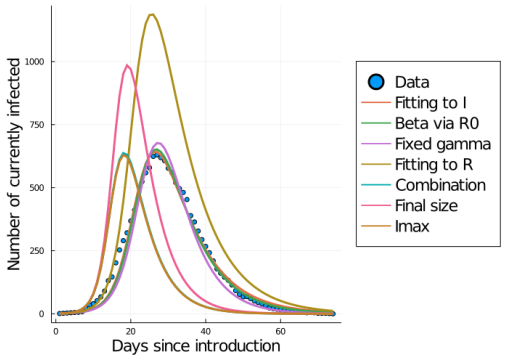
(a) Bridge-in (3.4).
Box-plot results for N^*.



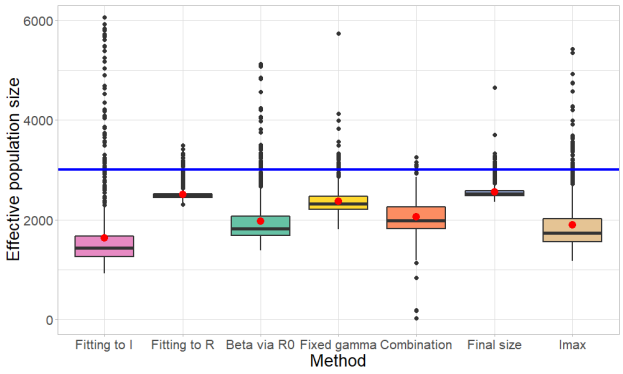
(b) Bridge-in (3.4).
Example of fitting results.



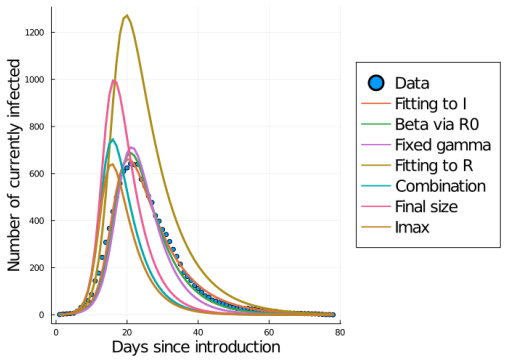
(c) Bridge-out (3.5).
Box-plot results for N^*.



(d) Bridge-out (3.5).
Example of fitting results.

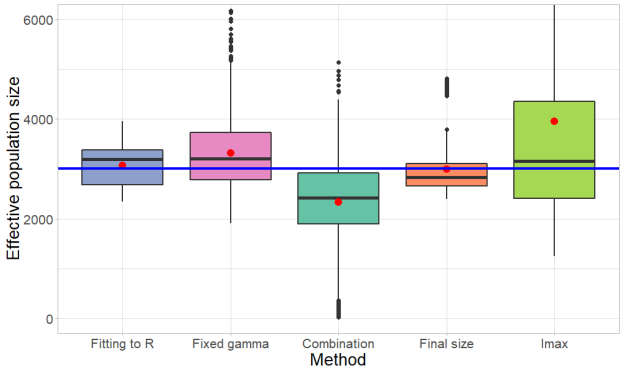


(e) Descending (3.6).
Box-plot results for N^*.

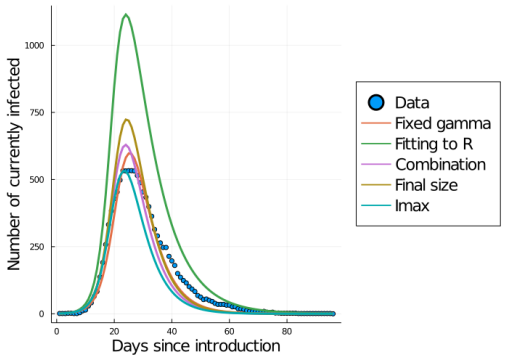


(f) Descending (3.6).
Example of fitting results.

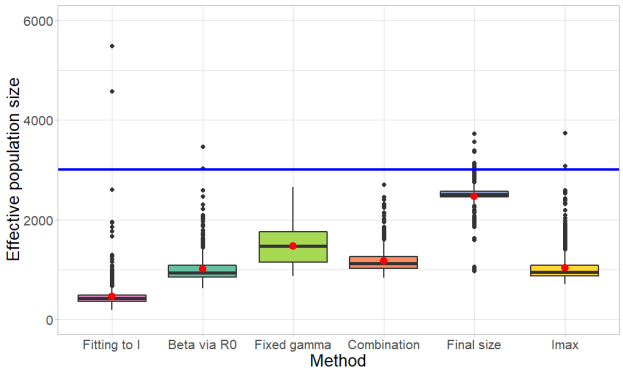
Figure 4.2: On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axis is restricted to $[0, 6000]$. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula.



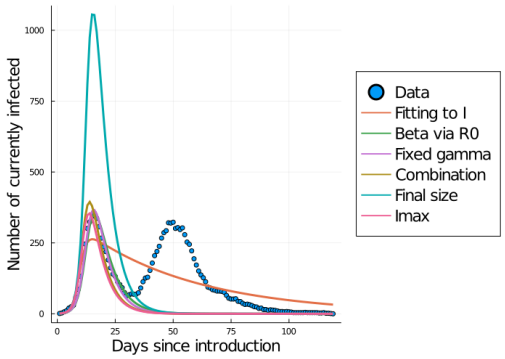
(a) Ascending (3.7).
Box-plot results for N^*.



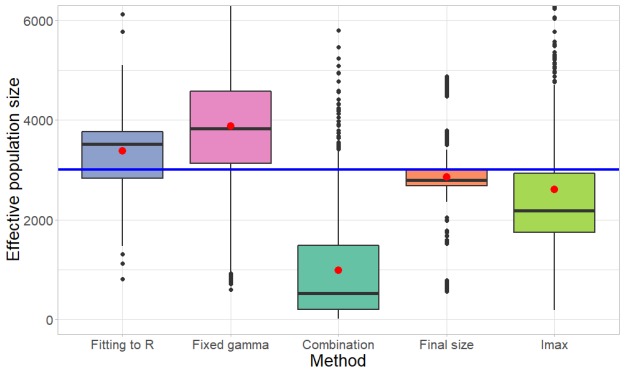
(b) Ascending (3.7).
Example of fitting results.



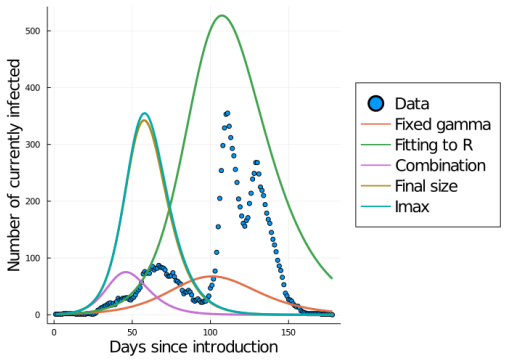
(c) Decreasing (3.8).
Box-plot results for N^*.



(d) Decreasing (3.8).
Example of fitting results.

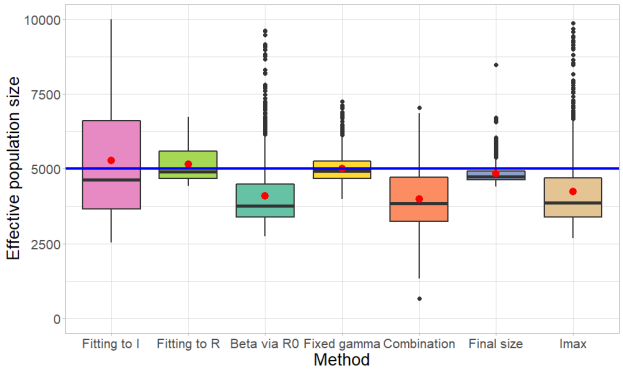


(e) Increasing (3.9).
Box-plot results for N^*.

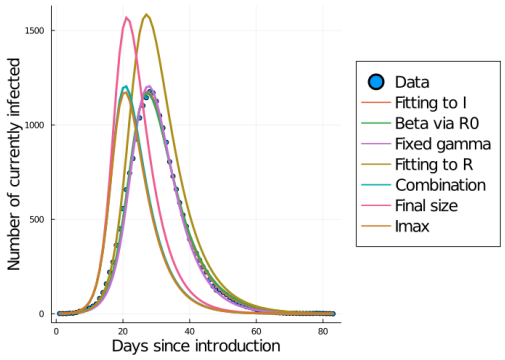


(f) Increasing (3.9).
Example of fitting results.

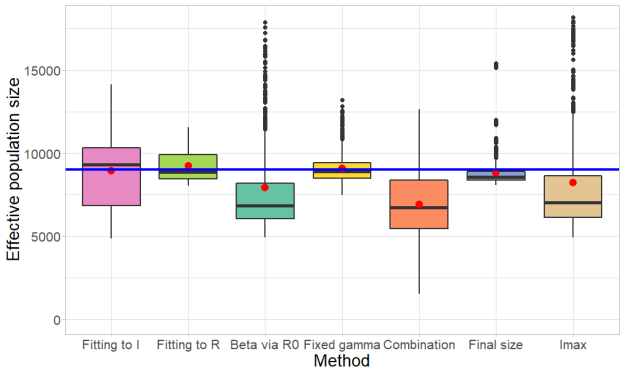
Figure 4.3: On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axis is restricted to $[0, 6000]$. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula.



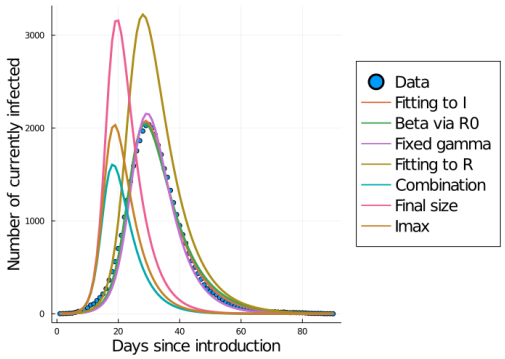
(a) Basic on 5 pacthes (5.1). Box-plot results for N^*.



(b) Basic on 5 patches (3.1). Example of fitting results.



(c) Basic on 9 patches (9.1). Box-plot results for N^*.



(d) Basic on 9 patches (9.1). Example of fitting results.

Figure 4.4: On the left: box-plots for optimum values of N^* obtained from the discussed methods applied to complex simulated outbreaks data. y-axes are restricted to $[0, 10000]$ and $[0, 18000]$ respectively. The blue line denotes the sum of population sizes in patches used for simulation N. The red dot denotes the mean. On the right: curves are fitted using optimum parameter values computed by methods. If γ is not computed, it is fixed at 0.2. If β is not computed, it is parametrized by $\mathcal{R}_0$ formula.

Chapter 5

Conclusions

5.1 Final discussion

In the previous 3 sections, we considered 3 applications of the methods derived in chapter 2 based on outbreak data types. These applications were a simple model simulation, a complex model simulation and real data. We had 2 main focuses: a comparison of N^* with N and the fitting potential of the standard model in system (1.4) based on computed optimum parameter values.

Methods that computed only N^* such as transcendental equations and combination did not provide adequate fitting to data. I_{max} and combination methods were not consistent in the computation of N^* in a homogeneous mixing setting. The final size method, although performing well on simple model-simulated data by getting $N^* \approx N$, was not found to be applicable in real outbreak data due to discrepancy between the initial growth rate and the final size of an outbreak. Similarly to the final size method, the fitting to R method performed well on simple model-simulated data, however in other 2 applications, sensitivity to an initial point for parameters, poor fitting to I-data and large confidence intervals led us to conclude that this method might not always be appropriate.

Based on simple simulated data results, fitting to I and β via $\mathcal{R}_0$ were found to be inadequate in computing N^* in homogeneous settings, despite having low corresponding SSE and AICs and thus appropriate for fitting purposes. Note that in some settings of simulated data, we were not able to apply them. The main reason was the absence of a starting point for parameter values that would work for all data and/or the data had too many extrema making the fitting algorithms fail. On the other hand, we could apply the fixed γ method in any derived setting for simulated data. Moreover, unlike other parameters such as β or $\mathcal{R}_0$, γ is often confidently known for emerging disease and does not majorly vary across locations. Optimum parameters computed by this method provided a decent fit to real outbreak data. Based on simulated data tests and fittings to real data, we viewed the fixed γ method as preferable compared to the rest. As can be seen in figure 5.1, depending on the target and availability of information, other methods may be more appropriate. We

did not include other criteria such as computational costs as they were not the primary goals of this thesis.

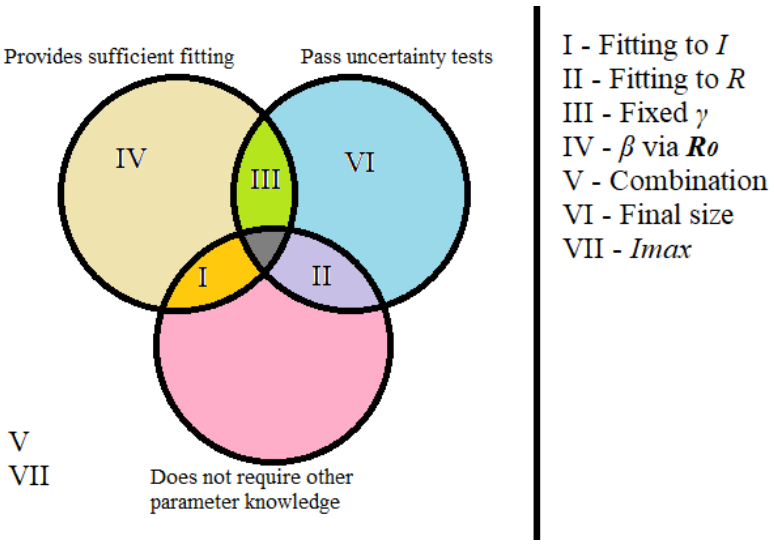


Figure 5.1: Summary on methods performance based on their application in chapters 2, 3 and 4. The details (e.g. what data and parameters are needed) on each of the methods can be found in section 2.1 and table 2.1.

The complex model simulations shed a light on the inadequacy of our methods in some settings. Since standard model (1.4) can capture only one bump in I-data, neither of our methods provided good fitting in some complex model simulations. Those settings were the ones where subpopulations were mostly isolated. Moreover, unlike the effective population size in genetics N_e, some methods in certain settings of simulated data provided N^*'s that were considerably larger than the population size used for these simulations. One of the possible reasons is that data suggested a slow initial growth of cases, thus a method would provide smaller β, however, there was a rapid increase in cases later, thus a method had to provide larger N^* to compensate for that. Nevertheless, in the case of real outbreak data, we saw that for any city and any method, N^* is clearly less than the census population size. Since we were confident in having $N^*/N < 1$ in case of real outbreak data and we had additional information (distance to Wuhan), we further analysed implications of this ratio with N^* computed by the fixed γ method in Chinese cities' COVID-19 data.

N^*/N solely could be used to assess how successfully an outbreak was contained. It could serve as a score, and one would expect N^*/N to be small if an outbreak was contained well. Correlations of N^*/N to other factors such as distances between locations allowed us to look at multiple cities. For example, in figures 3.13 and A.10 we expected locations to lie on a hyperbolic curve since there was a national lockdown implemented to all cities. Points that were not within a defined band (e.g. a location with relatively high N^*/N and distance from Wuhan) deserve a deeper investigation. The examples were Haerbin, the farthest city

from Wuhan (among selected ones), that had a medium value of N^*/N (among its cluster) and Ezhou, one of the closest cities to Wuhan, that had a relatively small value of N^*/N (among its cluster). Again, work needs to be done to define what should be considered as a deviating point (i.e. determine a band, range etc). The other factors that could be used for the correlation computation are time period before lockdown and density. We were not able to find necessary data on these factors, but this fact underlined the importance of the availability of such data.

5.2 Summary and Future work

We have introduced a new notion of the effective population size in epidemiology. In this thesis, the effective population size N^* is defined as the number of individuals necessary to get the same SIR dynamics as suggested by outbreak data. Our goals were to find N^* in various data types, to compare it to N (the true population size used for simulation or the census population size of a location for a real outbreak) and to observe corresponding curve fittings to data. We provided the necessary background and elaborated on the main definition with motivation in chapter 1. In chapter 2, we derived methods coming from model (1.4) to estimate optimum N^* and other relevant parameters. Then we applied those methods to simple and complex model simulated outbreak data (chapter 2 and 4) and real outbreak data (chapter 3).

In simple model simulations, we verified that some methods in controlled homogeneous settings would return $N^* \approx N$, as expected. When we applied the methods to data of COVID-19 in Chinese cities, we saw that N^* was considerably less than the corresponding census population size. At the same time, some methods returned optimum parameter values (including N^*) that provided a reasonable fit to real outbreak data, in our case, of COVID-19 in Chinese cities at the beginning of the pandemic. In general, treating a population size as unknown allowed us to fit a simple model such as system (1.4). One of the methods that both returned $N^* \approx N$ in simple model simulations and provided a decent fit on real outbreak data was the fixed γ method. We used optimum N^* computed by this method on COVID-19 outbreaks in Chinese cities to find correlations with other factors such as a distance between a given city and Wuhan. We found that N^*/N was reciprocal to this distance.

In our work, we focused on methods that could be derived from simple models. Such methodology could be less resource-consuming and more interpretable in comparison with others, for example, those with population structure. However, there are clear drawbacks and possible extensions to our methodology. First, using a simple standard model for method derivations limited the scope of data types that we could analyse as the methodology is more appropriate when applied to I-data with 1 maximum. For example, we saw that in some settings of complex simulated data, none of our methods succeeded in providing a

decent fit to data, as fitted curves were capturing either 1 of multiple maxima or going through their averages. However, note that this could rather be a limitation of the fitting method. Second, none of our methods returned the same N^* when applied to real data. This suggests that one needs to be careful in claiming estimates and should add a used method, e.g. instead of just saying N^* say N^* of fitting to I (method). Such ambiguity was also found in the effective population size in genetics (e.g. variance N_e and inbreeding N_e). Third, some methods required pre-estimated parameter values such as γ and the growth rate for $\mathcal{R}_0$. This might not be known in outbreaks of an emerging disease. Moreover, we applied our methods to data where I reached 0. Hence, our methods were appropriate in retrospective analysis although some methods (e.g. fixed γ) could be performed in partial data sets. Some of the listed drawbacks could be eliminated.

Our derived methodology can be improved in several directions. First, one needs to investigate how to deal with outbreak data that has multiple "bumps" (similar to complex model-simulated data in some settings). One could extend the number of compartments or dissect data appropriately, however, it is difficult to state which method is the optimum one and if interpretability (one of our main features) will not be lost. Second, in the case of complex simulation, we applied the methods to a limited number of settings. One could look at other settings (increasing the number of patches would increase the amount of population structures) and in general other heterogeneous model simulations (e.g. diffusion models). We expect that some of our conclusions will persist in higher dimensions. Third, it is desirable to have information on other location statistics and features such as density when considering real outbreaks. By computing relations between N^* or N^*/N and other factors, we could explore the potential of N^* more profoundly. One could compare which factors (e.g. density, distance to the index case, length of lockdown) are the most (positively/negatively) correlated with N^*/N.

If we turn to the utility of N^*, correlations of N^*/N and other factors (distance, time length, density etc.) could show which locations would deserve a deeper investigation in how an outbreak proceeded. Such (retrospective) analysis could be useful in informing public health teams with disease containing management directions. For example, lockdown is one of the more restrictive strategies (compared to social distancing), and N^* could be a tool in assessing its effectiveness. From the theoretical point of view, N^* is a parameter that we do not have a standard way of computing (as far as we know) and establishing one would be beneficial to the epidemiological community. In this thesis, we concluded the fixed γ method is one of the candidates. Nevertheless, we observed that the fitting to I method produced less spread results as N increased in simulations, and one could ask if there was such N that fitting to I would be as accurate as fixed γ. Theoretical questions such as the previous one would help to determine the best methods. Although some drawbacks cannot be eliminated without overhauling system (1.4) and corresponding methods, we believe that similarly to the effective population size in genomics N_e, the effective population size in epidemiology

N^* can be a major source of future research both in its implications on the outbreak and its accurate and relevant estimation.

Bibliography

[1] S. C. Anderson, A. M. Edwards, M. Yerlanov, N. Mulberry, J. E. Stockdale, S. A. Iyaniwura, R. C. Falcao, M. C. Otterstatter, M. A. Irvine, N. Z. Janjua, D. Coombs, and C. Colijn, *Quantifying the impact of covid-19 control measures using a bayesian model of physical distancing*, PLoS Computational Biology, 16 (2020).

[2] V. Andreasen, *Dynamics of annual influenza a epidemics with immuno-selection*, Journal of mathematical biology, 46 (2003), pp. 504–536.

[3] J. Arino, F. Brauer, P. van den Driessche, J. Watmough, and J. Wu, *Simple models for containment of a pandemic*, Journal of the Royal Society interface, 3 (2006), pp. 453–457.

[4] D. Bernoulli and S. Blower, *An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it*, Reviews in medical virology, 14 (2004), pp. 275–288.

[5] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A fresh approach to numerical computing*, SIAM review, 59 (2017), pp. 65–98. Packages: LsqFit, Optim, Plots, DifferentialEquations, DataFrames, CSV, LinearAlgebra, Roots, StatsBase.

[6] F. Brauer and C. Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*, vol. 40 of Texts in applied mathematics, Springer New York, New York, 2012.

[7] F. Brauer, C. Castillo-Chavez, and Z. Feng, *Mathematical Models in Epidemiology*, Springer, 2019.

[8] K. P. Burnham and D. R. Anderson, *Multimodel inference: Understanding aic and bic in model selection*, Sociological methods and research, 33 (2004), pp. 261–304.

[9] B. Charlesworth, *Effective population size and patterns of molecular evolution and variation*, Nature reviews. Genetics, 10 (2009), pp. 195–205.

[10] C. Darwin, *On the origin of species / by Charles Darwin.*, First Avenue Editions, a Division of Lerner Publishing Group, 2018.

[11] J. de Leeuw, *Information Theory and an Extension of the Maximum Likelihood Principle by Hirotogu Akaike*, eScholarship, University of California, 2011.

[12] C. A. Donnelly, A. C. Ghani, G. M. Leung, A. J. Hedley, C. Fraser, S. Riley, L. J. Abu-Raddad, L.-M. Ho, T.-Q. Thach, P. Chau, K.-P. Chan, T.-H. Lam, L.-Y. Tse, T. Tsang, S.-H. Liu, J. H. Kong, E. M. Lau, N. M. Ferguson, and R. M. Anderson, *Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in hong kong*, The Lancet (British edition), 361 (2003), pp. 1761–1766.

[13] Z. Du, X. Xu, Y. Wu, L. Wang, B. Cowling, and L. Meyers, *Covid-19 serial interval estimates based on confirmed cases in public reports from 86 chinese cities*, medRxiv, (2020). Preprint.

[14] L. R. Elveback, J. P. Fox, E. Ackerman, A. Langworthy, M. Boyd, and L. Gatewood, *An influenza simulation model for immunization studies*, American journal of epidemiology, 103 (1976), pp. 152–165.

[15] J. Hawks, *From genes to numbers: Effective population sizes in human evolution*, in Recent Advances in Palaeodemography, Springer Netherlands, Dordrecht, 2008, pp. 9–30.

[16] C. M. Hurvich and C.-L. Tsai, *Regression and time series model selection in small samples*, Biometrika, 76 (1989), pp. 297–307.

[17] J. A. Jacquez and P. O'Neill, *Reproduction numbers and thresholds in stochastic epidemic models i. homogeneous populations*, Mathematical biosciences., 107 (1991), pp. 161–186.

[18] D. Kang, H. Choi, J.-H. Kim, and J. Choi, *Spatial epidemic dynamics of the covid-19 outbreak in china*, International journal of infectious diseases, 94 (2020), pp. 96–102.

[19] W. O. Kermack and A. G. McKendrick, *Contributions to the mathematical theory of epidemics—i*, Bulletin of mathematical biology, 53 (1991), pp. 33–55.

[20] ——, *Contributions to the mathematical theory of epidemics—ii. the problem of endemicity*, Bulletin of mathematical biology, 53 (1991), pp. 57–87.

[21] ——, *Contributions to the mathematical theory of epidemics—iii. further studies of the problem of endemicity*, Bulletin of mathematical biology, 53 (1991), pp. 89–118.

[22] W. R. KhudaBukhsh, B. Choi, E. Kenah, and G. A. Rempała, *Survival dynamical systems: individual-level survival analysis from population-level epidemic models*, Interface focus, 10 (2020), pp. 20190048–20190048.

[23] R. Kliman, B. Sheehy, and J. Schultz, *Genetic drift and effective population size*, Nature Education, 1 (2008), p. 3.

[24] C. D. Lab, *China COVID-19 Daily Cases with Basemap*, 2020.

[25] T. T.-Y. Lam, C.-C. Hon, and J. W. Tang, *Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections*, Critical reviews in clinical laboratory sciences, 47 (2010), pp. 5–49.

[26] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quarterly of applied mathematics, 2 (1944), pp. 164–168.

[27] B. Li and B. Lu, *How china made its covid-19 lockdown work*, 2020. Assessed on 22.06.2021, https://www.eastasiaforum.org/2020/04/07/how-china-made-its-covid-19-lockdown-work/.

[28] J. Li, D. Blakeley, and R. J. Smith?, *The failure of r0*, Computational and mathematical methods in medicine, 2011 (2011), pp. 527610–527610.

[29] M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray, *Transmission dynamics and control of severe acute respiratory syndrome*, Science, 300 (2003), pp. 1966–1970.

[30] D. W. Marquardt, *An algorithm for least-squares estimation of nonlinear parameters*, Journal of the Society for Industrial and Applied Mathematics, 11 (1963), pp. 431–441.

[31] M. Martcheva, *An Introduction to Mathematical Biology*, Springer, 2015.

[32] Microsoft, *Visual Studio Code*. `https://code.visualstudio.com/`.

[33] S. Offner, *Mendel's peas and the nature of the gene: Genes code for proteins and proteins determine phenotype*, The American Biology Teacher, 73 (2011), pp. 382 – 387.

[34] W. H. Organization, *Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020*. Assessed on 22.03.2021, `https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020`.

[35] P. O'Neill and T. Kypraios, *Mcmc 2 for infectious diseases*. Accessed on 11.12.2020, `https://www.maths.nottingham.ac.uk/plp/pmztk/files/MCMC2-Seattle/`.

[36] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. Packages: tidy-verse, ggmap, forcats, maps, maptools.

[37] G. F. Raggett, *Modelling the eyam plague*, Institute of Mathematics and its Applications, 18 (1982), pp. 221–226.

[38] M. Roberts, *The pluses and minuses of r0*, Journal of the Royal Society interface, 4 (2007), pp. 949–961.

[39] R. Ross, *The Prevention of Malaria*, John Murray, London, 1911.

[40] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. `http://www.rstudio.com/`.

[41] L. Sattenspiel and K. Dietz, *A Structured Epidemic Model Incorporating Geographic*, Mathematical Biosciences, 91 (1994), pp. 71–91.

[42] F. Standl, K.-H. Jöckel, B. Brune, B. Schmidt, and A. Stang, *Comparing sars-cov-2 with sars-cov and influenza pandemics*, The Lancet infectious diseases, 21 (2021), pp. e77–e77.

[43] H. R. Thieme and C. Castillo-Chavez, *How may infection-age-dependent infectivity affect the dynamics of hiv/aids?*, SIAM journal on applied mathematics, 53 (1993), pp. 1447–1479.

[44] J. Wang, E. Santiago, and A. Caballero, *Prediction and estimation of effective population size*, Heredity, 117 (2016), pp. 193–206.

[45] J. N. Weitzel, K. R. Blazer, D. J. MacDonald, J. O. Culver, and K. Offit, *Genetics, genomics, and cancer risk assessment: State of the art and future directions in the era of personalized medicine*, CA: a cancer journal for clinicians, 61 (2011), pp. 327–n/a.

[46] Wikipedia, *Effective population size*. Accessed on 18.01.2021, `https://en.wikipedia.org/wiki/Effective_population_size#cite_note-1`.

[47] J. W. Wood, *The genetic demography of the gainj of papua new guinea. 2. determinants of effective population size*, The American naturalist, 129 (1987), pp. 165–187.

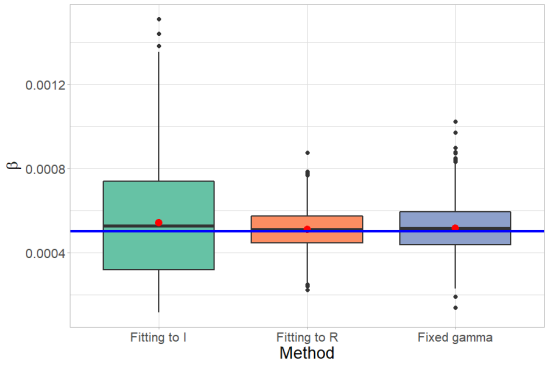
[48] J. Wu, van den Driessche Pauline, and F. Brauer, *Mathematical Epidemiology*, Springer Berlin / Heidelberg, Berlin, Heidelberg, 2008.

[49] J. T. Wu, K. Leung, and G. M. Leung, *Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study*, The Lancet (British edition), 395 (2020), pp. 689–697.

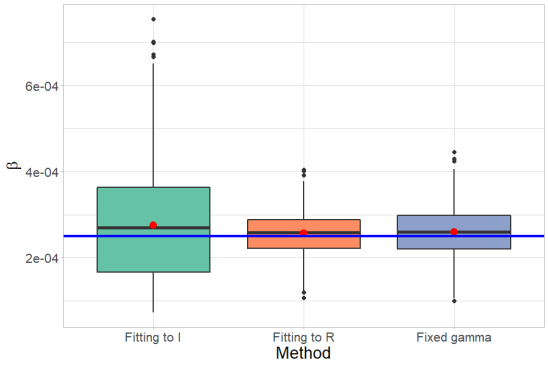
[50] S. Zhang, Z. Wang, R. Chang, H. Wang, C. Xu, X. Yu, L. Tsamlag, Y. Dong, H. Wang, and Y. Cai, *Covid-19 containment: China provides important lessons for global response*, Frontiers of medicine, 14 (2020), pp. 215–219.

Appendix A

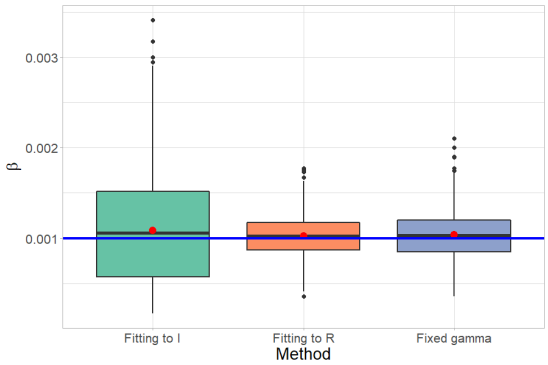
Supplemental figures



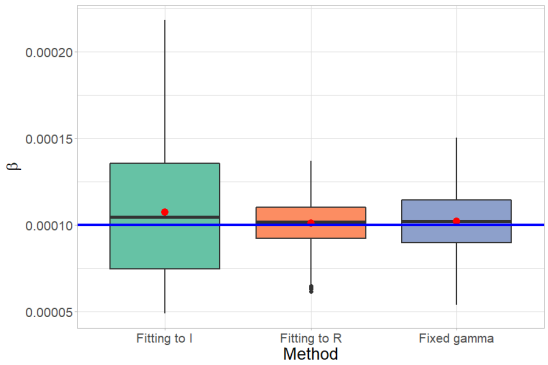
(a) The baseline setting.
$N = 1000$, $\beta = 0.0005$, $\gamma = 0.2$



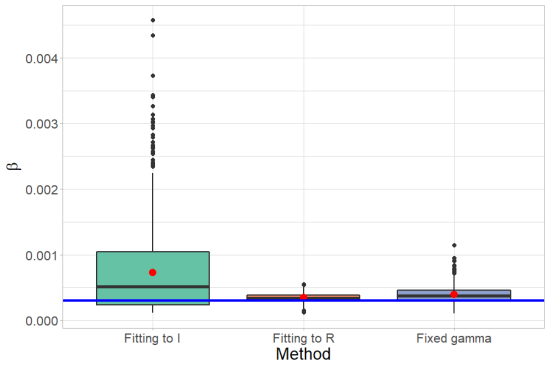
(b) Modified rates.
$N = 1000$, $\beta = 0.00025$, $\gamma = 0.1$



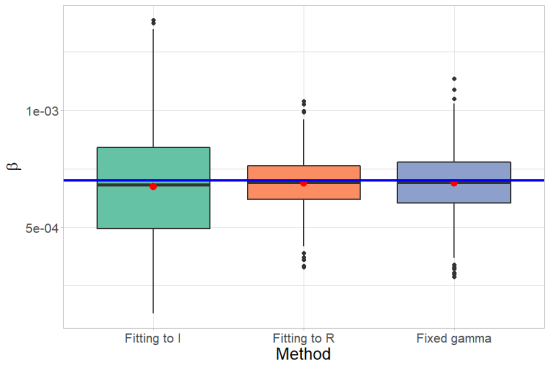
(c) Smaller population.
$N = 500$, $\beta = 0.001$, $\gamma = 0.2$



(d) Larger population.
$N = 5000$, $\beta = 0.0001$, $\gamma = 0.2$

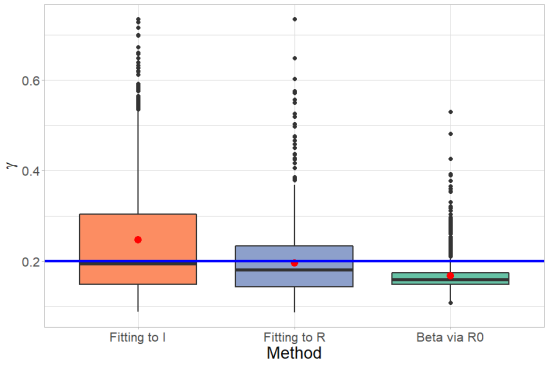


(e) Smaller $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0003$, $\gamma = 0.2$

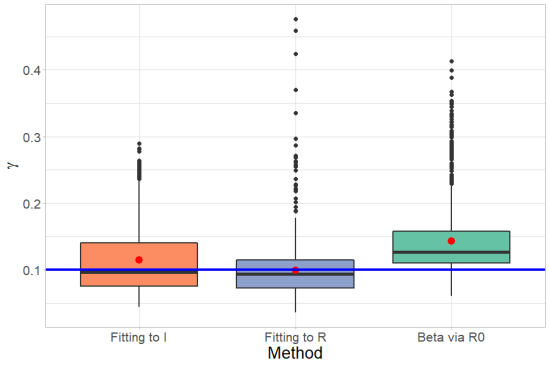


(f) Larger $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0007$, $\gamma = 0.2$

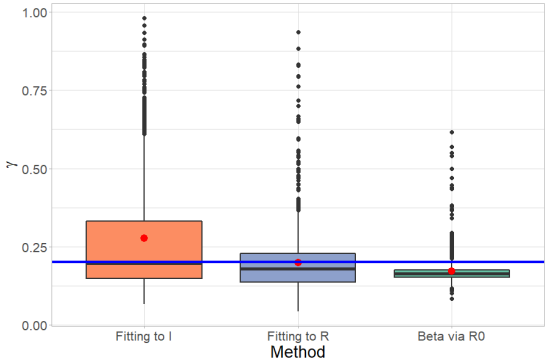
Figure A.1: Box-plots for values of β (y-axis) obtained from the discussed methods. The blue line denotes the true value. The red dot denotes the mean.



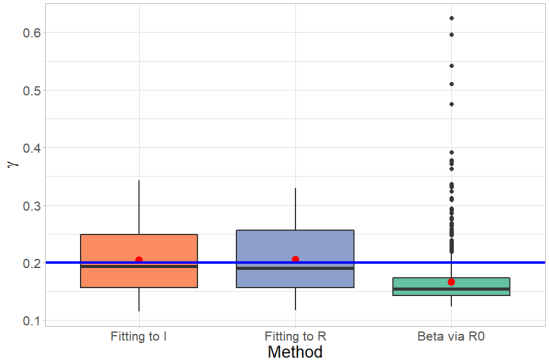
(a) The baseline setting.
$N = 1000$, $\beta = 0.0005$, $\gamma = 0.2$



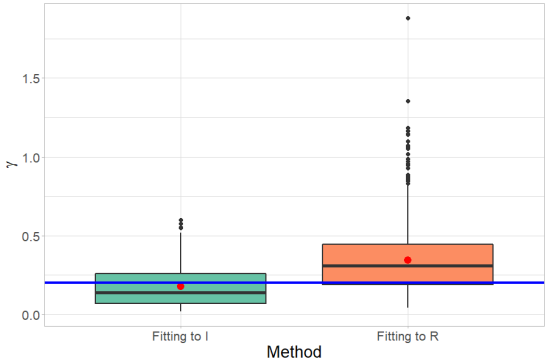
(b) Modified rates.
$N = 1000$, $\beta = 0.00025$, $\gamma = 0.1$



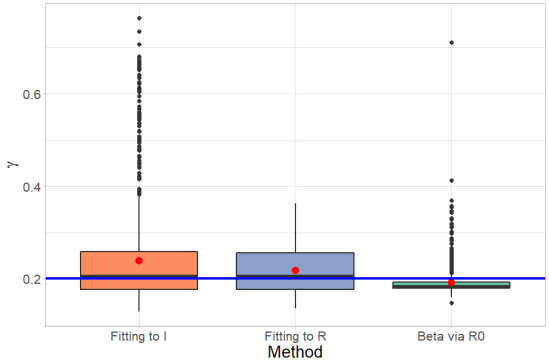
(c) Smaller population.
$N = 500$, $\beta = 0.001$, $\gamma = 0.2$



(d) Larger population.
$N = 5000$, $\beta = 0.0001$, $\gamma = 0.2$

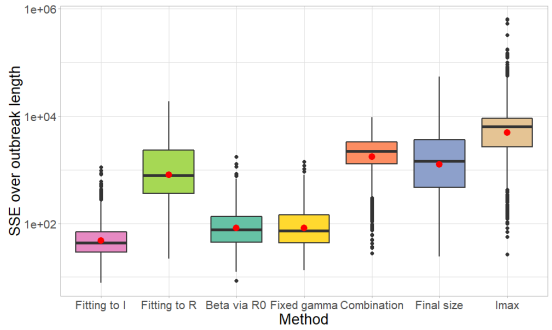


(e) Smaller $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0003$, $\gamma = 0.2$

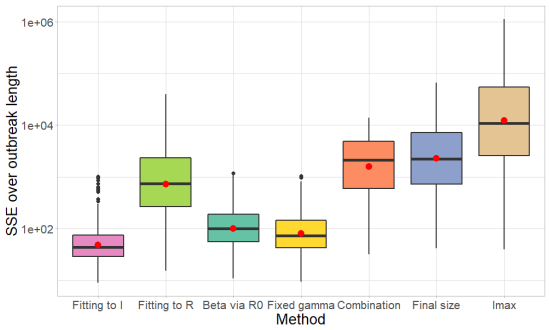


(f) Larger $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0007$, $\gamma = 0.2$

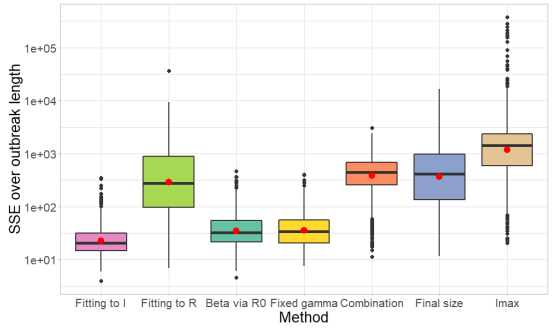
Figure A.2: Box-plots for values of γ (y-axis) obtained from some of the discussed methods. The blue line denotes the true value. The red dot denotes the mean.



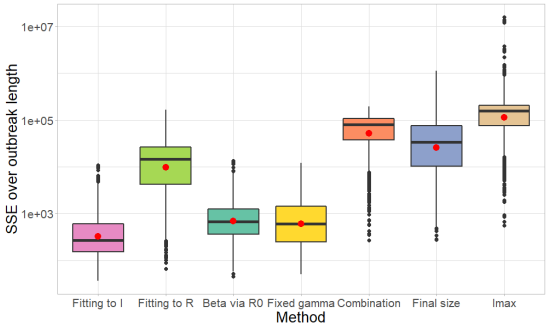
(a) The standard setting.
$N = 1000$, $\beta = 0.0005$, $\gamma = 0.2$



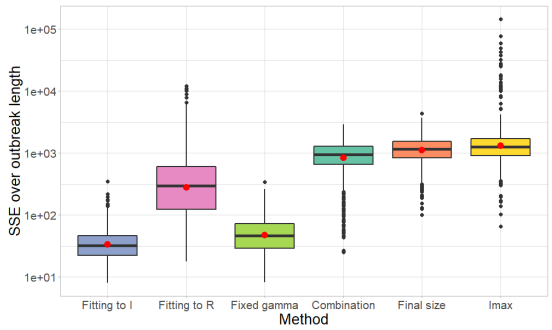
(b) Modified rates.
$N = 1000$, $\beta = 0.00025$, $\gamma = 0.1$



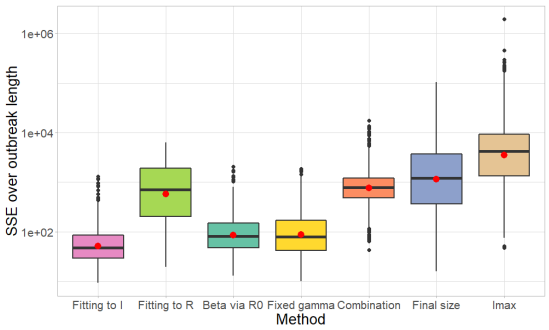
(c) Smaller population.
$N = 500$, $\beta = 0.001$, $\gamma = 0.2$



(d) Larger population.
$N = 5000$, $\beta = 0.0001$, $\gamma = 0.2$

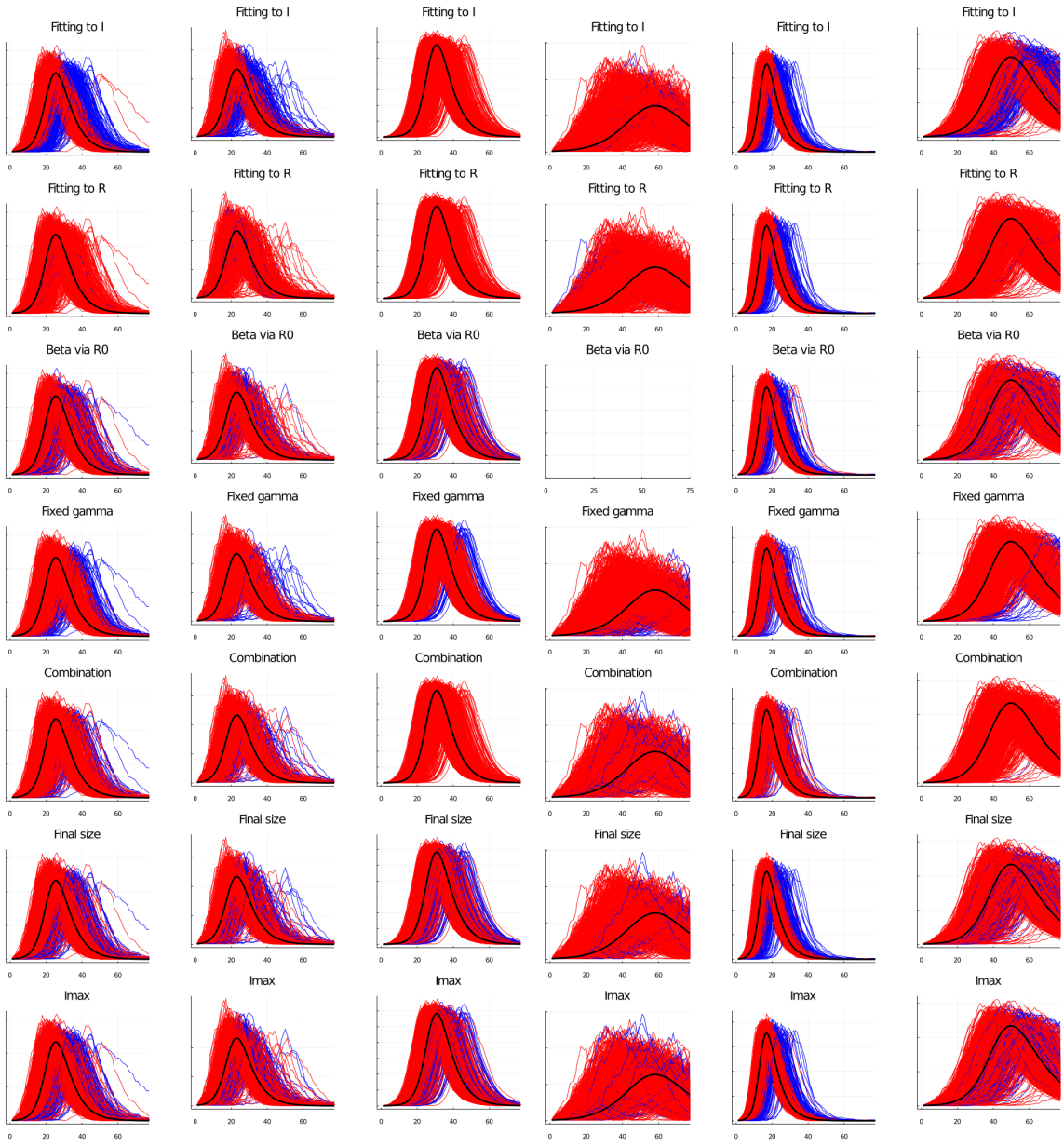


(e) Smaller $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0003$, $\gamma = 0.2$



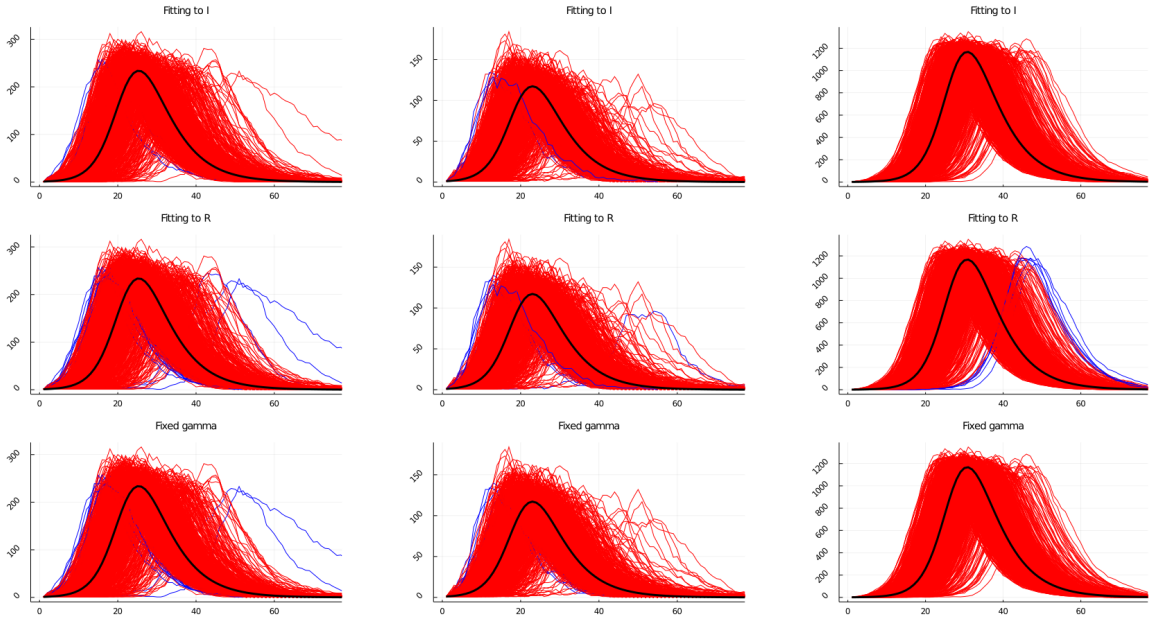
(f) Larger $\mathcal{R}_0$.
$N = 1000$, $\beta = 0.0007$, $\gamma = 0.2$

Figure A.3: Box-plots for values of average SSE over length of an outbreak (y-axis) obtained from the discussed methods. y-axis is on log-scale. Parameters that could not be computed using a method (such as transcendental equations) are taken as true ones. The red dot denotes the mean.



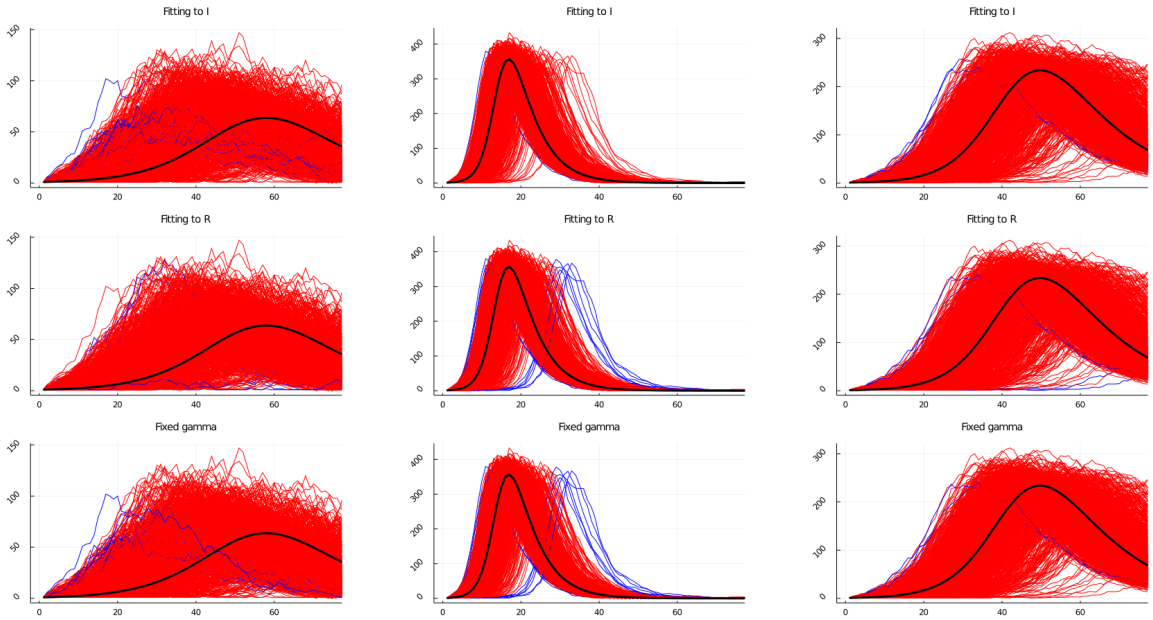
(a) Baseline (b) Smaller population (c) Larger population (d) Smaller $\mathcal{R}_0$ (e) Larger $\mathcal{R}_0$ (f) Modified rates

Figure A.4: N^*-outliers in data across all discussed methods and all settings. Blue — outliers, red — non-outliers. y-axis is the number of individuals.



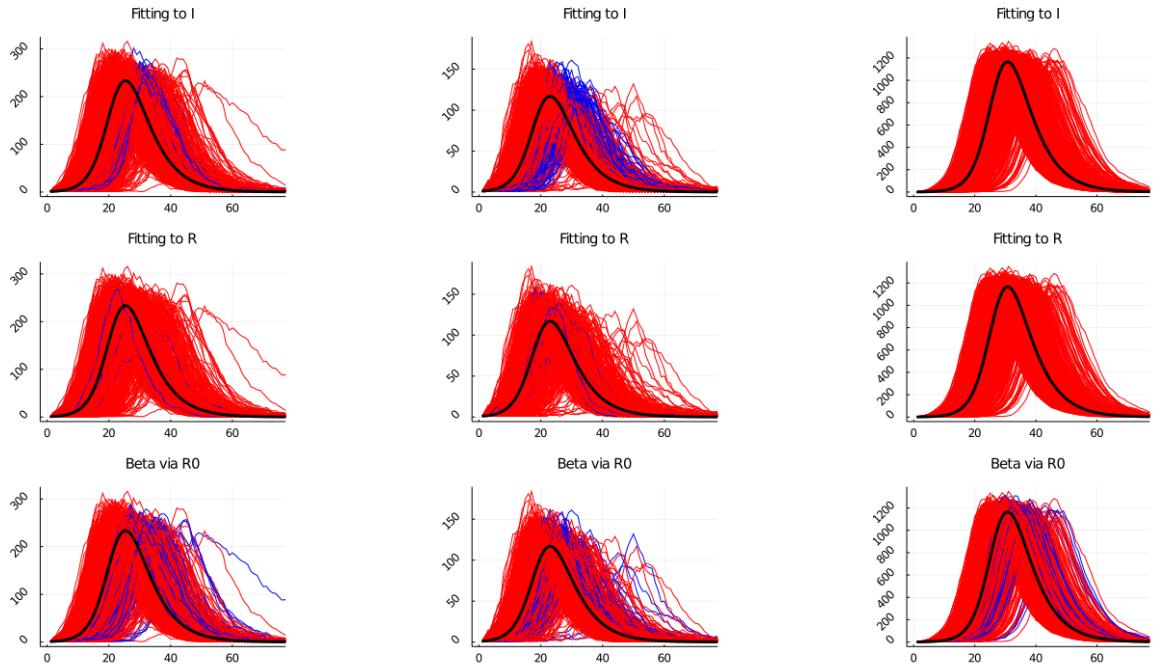
(a) Baseline (b) Smaller population (c) Larger population

Figure A.5: β-outliers in data across 3 discussed methods and all settings. Blue — outliers, red — non-outliers. y-axis is the number of individuals.



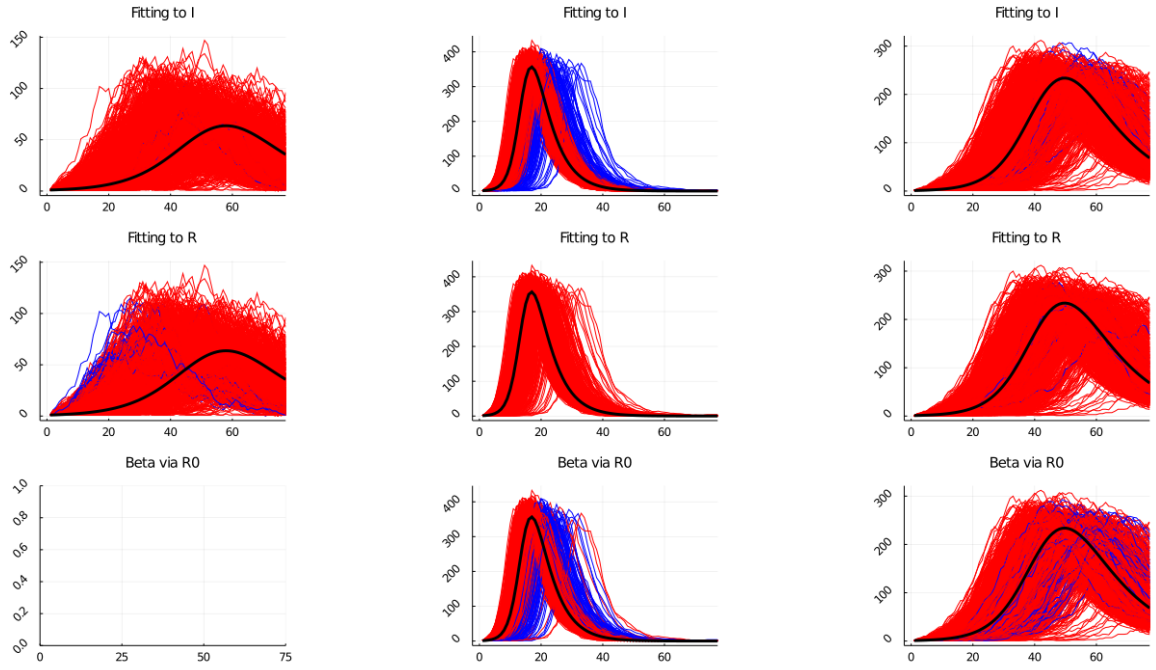
(a) Smaller $\mathcal{R}_0$ (b) Larger $\mathcal{R}_0$ (c) Modified rates

Figure A.6: β-outliers in data across 3 discussed methods and all settings. Blue — outliers, red — non-outliers. y-axis is the number of individuals.



(a) Baseline (b) Smaller population (c) Larger population

Figure A.7: γ-outliers in data across 3 discussed methods and all settings. Blue — outliers, red — non-outliers. y-axis is the number of individuals.



(a) Smaller $\mathcal{R}_0$ (b) Larger $\mathcal{R}_0$ (c) Modified rates

Figure A.8: γ-outliers in data across 3 discussed methods and all settings. Blue — outliers, red — non-outliers. y-axis is the number of individuals.

City name	Fitting to I, Lower CI	Fitting to I	Fitting to I, Upper CI	β via $\mathcal{R}_0$, Lower CI	β via $\mathcal{R}_0$	β via $\mathcal{R}_0$, Upper CI
Tianjin	251	328	405	300	313	326
Haerbin	303	355	407	508	539	570
Nanjing	134	184	234	196	206	217
Xuzhou	128	182	236	166	177	187
Suzhou	121	136	150	233	249	266
Huaiyin	106	129	153	170	178	187
Hangzhou	232	255	278	384	412	440
Ningbo	368	425	483	415	427	439
Wenzhou	652	748	844	990	1068	1147
Taizhou	170	186	203	326	361	395
Hefei	256	305	353	403	428	453
Bengbu	488	703	918	488	510	531
Anqing	128	158	188	209	224	239
Fuyang	251	328	406	357	378	400
Liuan	203	270	337	232	238	245
Bozhou	245	320	395	296	310	323
Fuzhou	90	113	136	178	198	218
Putian	60	70	79	150	172	194
Nanchang	282	329	377	452	486	520
Jiujiang	155	181	207	302	332	362
Xinyu	206	256	306	297	315	332
Ganzhou	145	196	246	211	224	237
Yichun	166	310	455	251	271	292
Fuzhou	182	227	272	234	242	249
Shangrao	337	435	533	331	343	356
Zhengzhou	153	180	207	310	342	375
Nanyang	258	379	499	309	330	351
Shangqiu	210	294	377	211	221	231
Xinyang	380	474	567	576	623	669
Zhoukou	82	96	109	177	200	222
Zhumadian	255	304	354	324	337	351
Huangshi	1976	2203	2431	1768	1823	1877
Shiyan	735	849	962	1279	1409	1539
Yichang	1427	1614	1801	1744	1832	1920
Xiangfan	1573	1775	1977	2445	2665	2886
Ezhou	1670	1904	2138	2514	2707	2899
Jingmen	1176	1360	1544	1821	1970	2118
Xiaogan	5584	6528	7472	6148	6444	6739
Jingzhou	2342	2675	3008	2883	3023	3163
Huanggang	4227	4781	5336	4822	5014	5207
Xianning	1572	5120	8667	1179	1292	1406
Suizhou	1815	2017	2220	2645	2829	3012
Changsha	291	317	342	546	595	644
Zhuzhou	115	138	160	192	202	213
Shaoyang	157	183	210	220	230	241
Yueyang	221	253	285	338	355	373
Changde	97	121	144	173	188	203
Loudi	132	150	168	175	181	186
Zhuhai	181	220	258	290	302	315
Huizhou	149	235	320	171	178	186
Dongguan	149	200	252	244	263	282
Chongqing	716	812	909	1040	1105	1170
Xian	237	276	314	302	313	323

Table A.1: N^* computed using 2 methods for chosen 53 cities. Confidence intervals are 95%.

City name	Fitting to R, Lower CI	Fitting to R	Fitting to R, Upper CI	Fixed γ, Lower CI	Fixed γ	Fixed γ, Upper CI
Tianjin	144	190	237	376	390	404
Haerbin	194	266	338	595	629	662
Nanjing	68	278	488	289	303	318
Xuzhou	81	99	118	185	195	205
Suzhou	73	131	190	257	274	290
Huaiyin	51	97	143	165	171	178
Hangzhou	458	505	551	394	419	443
Ningbo	141	239	337	461	473	485
Wenzhou	645	848	1052	1269	1373	1478
Taizhou	200	265	331	366	402	438
Hefei	199	232	266	458	485	511
Bengbu	-11881	3995	19872	467	485	503
Anqing	74	131	188	205	218	230
Fuyang	-573	627	1826	437	461	485
Liuan	-14018	5531	25080	225	231	237
Bozhou	39	193	346	283	294	305
Fuzhou	46	134	222	181	197	214
Putian	-43641	7909	59458	152	171	189
Nanchang	231	278	326	500	535	569
Jiujiang	153	167	181	341	372	403
Xinyu	137	165	193	308	324	339
Ganzhou	-55	281	618	234	246	258
Yichun	306	328	351	265	283	301
Fuzhou	70	128	185	216	222	228
Shangrao	146	173	200	345	356	367
Zhengzhou	-26108	4109	34326	341	373	405
Nanyang	85	262	439	439	463	488
Shangqiu	-48536	6566	61668	250	259	269
Xinyang	216	367	518	638	684	730
Zhoukou	9	182	356	181	201	220
Zhumadian	-30877	6600	44077	331	342	354
Huangshi	1186	1245	1304	3270	3368	3467
Shiyan	1519	1652	1786	1663	1837	2011
Yichang	866	1128	1390	3179	3369	3558
Xiangfan	1234	1602	1970	3036	3317	3599
Ezhou	1538	1659	1781	3441	3740	4039
Jingmen	1895	1996	2096	2411	2617	2822
Xiaogan	2778	4445	6112	10853	11518	12184
Jingzhou	1612	1960	2308	4462	4719	4975
Huanggang	3258	3470	3681	7849	8280	8711
Xianning	445	1462	2479	2628	2776	2923
Suizhou	1401	1553	1704	3751	4058	4365
Changsha	312	341	371	619	672	724
Zhuzhou	88	95	103	201	211	221
Shaoyang	99	118	138	214	222	231
Yueyang	168	233	299	438	461	485
Changde	70	128	186	201	215	230
Loudi	68	96	125	209	215	221
Zhuhai	-28601	5965	40531	346	360	373
Huizhou	-32897	4964	42824	170	177	183
Dongguan	124	135	147	289	309	328
Chongqing	648	708	769	1512	1619	1726
Xian	115	140	166	370	382	394

Table A.2: N^* computed using 2 methods for chosen 53 cities. Confidence intervals are 95%.

City name	Combination, Lower CI	Combination	Combination, Upper CI	I_{max}	Final size
Tianjin	323	354	386	320	144
Haerbin	547	595	643	492	209
Nanjing	207	242	276	206	98
Xuzhou	178	189	200	174	82
Suzhou	249	267	285	225	92
Huaiyin	165	172	180	176	71
Hangzhou	394	418	443	375	186
Ningbo	439	461	484	407	164
Wenzhou	1047	1207	1366	915	512
Taizhou	354	391	428	313	150
Hefei	430	467	503	387	180
Bengbu	461	483	506	484	171
Anqing	205	218	231	208	87
Fuyang	380	424	467	355	161
Liuan	226	232	238	239	77
Bozhou	281	294	308	294	114
Fuzhou	181	198	215	175	76
Putian	153	172	191	154	59
Nanchang	483	522	560	414	237
Jiujiang	326	359	391	282	123
Xinyu	307	322	338	281	136
Ganzhou	225	240	254	210	80
Yichun	262	280	299	261	223
Fuzhou	212	222	233	229	79
Shangrao	342	354	366	333	130
Zhengzhou	333	365	398	328	164
Nanyang	341	389	437	317	160
Shangqiu	228	245	263	216	94
Xinyang	615	668	720	538	282
Zhoukou	182	202	221	183	80
Zhumadian	330	342	354	336	144
Huangshi	1532	1982	2431	1753	1025
Shiyan	1337	1569	1800	1174	683
Yichang	1633	2029	2426	1744	940
Xiangfan	2626	2974	3323	2182	1184
Ezhou	2342	2863	3384	2366	1412
Jingmen	1823	2146	2469	1888	941
Xiaogan	4523	6171	7819	5834	3543
Jingzhou	2510	3169	3828	2997	1601
Huanggang	3643	4957	6272	4842	2934
Xianning	1007	1403	1799	1565	845
Suizhou	2619	3154	3688	2450	1320
Changsha	592	649	707	500	250
Zhuzhou	199	209	219	208	85
Shaoyang	214	223	233	218	106
Yueyang	366	409	451	349	162
Changde	189	205	221	174	86
Loudi	190	203	215	188	80
Zhuhai	316	339	361	307	106
Huizhou	171	177	183	193	68
Dongguan	265	290	314	262	106
Chongqing	1041	1252	1463	1004	585
Xian	325	352	378	304	126

Table A.3: N^* computed using 3 methods for chosen 53 cities. Confidence intervals are 95%.

City name	Fitting to I, lower CI	Fitting to I	Fitting to I, upper CI	Fitting to R, lower CI	Fitting to R	Fitting to R, upper CI	Fixed γ, lower CI	Fixed γ	Fixed γ, upper CI
Tianjin	1.046	1.275	1.504	1.125	1.344	1.563	1.072	1.118	1.165
Haerbin	1.026	1.182	1.338	1.080	1.160	1.240	0.697	0.739	0.781
Nanjing	1.501	1.948	2.394	0.327	0.924	1.520	1.247	1.320	1.393
Xuzhou	2.064	2.740	3.417	2.323	3.077	3.831	2.430	2.591	2.752
Suzhou	2.968	3.328	3.687	1.568	2.224	2.880	1.657	1.777	1.896
Huaiyin	3.046	3.599	4.153	2.198	3.116	4.035	2.729	2.861	2.992
Hangzhou	2.105	2.328	2.551	1.312	1.425	1.538	1.362	1.457	1.552
Ningbo	1.018	1.141	1.264	0.882	1.156	1.430	1.020	1.050	1.079
Wenzhou	0.627	0.722	0.818	0.509	0.546	0.583	0.372	0.406	0.439
Taizhou	2.726	3.037	3.347	1.570	1.793	2.016	1.249	1.385	1.522
Hefei	1.358	1.591	1.824	1.130	1.339	1.548	1.007	1.072	1.136
Bengbu	0.636	0.805	0.974	-0.238	0.096	0.430	1.010	1.052	1.095
Anqing	2.815	3.419	4.023	1.646	2.412	3.177	2.409	2.574	2.740
Fuyang	1.120	1.402	1.684	-0.245	0.497	1.239	1.001	1.062	1.124
Liuan	1.382	1.672	1.961	-0.159	0.076	0.312	1.814	1.868	1.921
Bozhou	1.384	1.679	1.974	0.600	1.587	2.574	1.717	1.791	1.865
Fuzhou	3.724	4.745	5.766	1.048	2.053	3.057	2.477	2.741	3.005
Putian	7.464	9.103	10.742	-0.400	0.084	0.568	2.794	3.194	3.595
Nanchang	1.410	1.641	1.871	1.381	1.432	1.484	0.978	1.050	1.123
Jiujiang	2.395	2.841	3.286	1.439	1.635	1.831	1.268	1.394	1.520
Xinyu	1.667	1.999	2.330	1.972	2.109	2.246	1.556	1.642	1.729
Ganzhou	1.867	2.385	2.902	0.135	0.994	1.853	1.864	1.979	2.093
Yichun	1.100	1.680	2.261	0.903	1.037	1.171	1.677	1.806	1.935
Fuzhou	1.752	2.051	2.351	1.398	2.063	2.729	2.023	2.086	2.149
Shangrao	1.007	1.203	1.400	1.441	1.649	1.857	1.350	1.399	1.448
Zhengzhou	2.358	2.806	3.255	-0.811	0.151	1.113	1.275	1.407	1.539
Nanyang	0.854	1.169	1.484	0.745	0.964	1.183	0.923	0.988	1.054
Shangqiu	1.361	1.751	2.141	-0.543	0.089	0.721	1.845	1.935	2.026
Xinyang	0.942	1.148	1.355	0.996	1.078	1.159	0.767	0.828	0.888
Zhoukou	4.949	5.973	6.997	0.667	1.964	3.262	2.406	2.706	3.006
Zhumadian	1.570	1.822	2.074	-0.351	0.089	0.529	1.587	1.653	1.718
Huangshi	0.168	0.185	0.202	0.169	0.195	0.221	0.127	0.131	0.136
Shiyan	0.552	0.644	0.737	0.417	0.426	0.435	0.268	0.297	0.327
Yichang	0.226	0.257	0.288	0.180	0.217	0.253	0.123	0.132	0.140
Xiangfan	0.333	0.380	0.427	0.283	0.295	0.307	0.176	0.193	0.210
Ezhou	0.233	0.266	0.298	0.223	0.236	0.249	0.133	0.145	0.157
Jingmen	0.331	0.383	0.434	0.327	0.332	0.338	0.193	0.210	0.227
Xiaogan	0.057	0.066	0.075	0.060	0.068	0.076	0.038	0.041	0.044
Jingzhou	0.148	0.167	0.187	0.164	0.170	0.177	0.097	0.104	0.110
Huanggang	0.084	0.094	0.104	0.090	0.092	0.095	0.056	0.059	0.063
Xianning	0.051	0.095	0.138	0.117	0.156	0.195	0.134	0.143	0.153
Suizhou	0.235	0.263	0.292	0.206	0.211	0.216	0.123	0.134	0.145
Changsha	1.490	1.632	1.774	1.023	1.085	1.146	0.731	0.798	0.865
Zhuzhou	2.735	3.200	3.666	3.094	3.257	3.420	2.139	2.255	2.372
Shaoyang	2.685	3.101	3.517	2.550	3.117	3.683	2.491	2.612	2.733
Yueyang	1.475	1.682	1.890	1.090	1.271	1.453	0.939	0.997	1.054
Changde	3.163	4.011	4.858	1.817	2.236	2.654	2.095	2.284	2.474
Loudi	2.463	2.754	3.046	2.575	2.939	3.303	1.993	2.064	2.135
Zhuhai	1.415	1.662	1.910	-0.300	0.075	0.450	1.086	1.137	1.187
Huizhou	1.572	2.094	2.615	-0.531	0.093	0.718	2.452	2.558	2.665
Dongguan	1.638	2.106	2.573	1.716	1.914	2.111	1.372	1.474	1.577
Chongqing	0.503	0.574	0.644	0.391	0.417	0.443	0.283	0.305	0.328
Xian	1.373	1.556	1.739	1.615	1.948	2.281	1.158	1.202	1.245

Table A.4: β computed using 3 methods for chosen 53 cities. Values are in 10^{-3}. Confidence intervals are 95%.

City name	Fitting to I, lower CI	Fitting to I	Fitting to I, upper CI	Fitting to R, lower CI	Fitting to R	Fitting to R, upper CI	β via $\mathcal{R}_0$, lower CI	β via $\mathcal{R}_0$	β via $\mathcal{R}_0$, upper CI
Tianjin	0.110	0.142	0.174	-0.025	0.058	0.140	0.132	0.135	0.139
Haerbin	0.074	0.090	0.105	0.031	0.141	0.252	0.139	0.143	0.147
Nanjing	0.074	0.103	0.133	0.113	0.178	0.242	0.112	0.116	0.120
Xuzhou	0.110	0.156	0.202	-0.064	0.072	0.207	0.145	0.151	0.158
Suzhou	0.063	0.073	0.083	0.000	0.037	0.073	0.146	0.151	0.157
Huaiyin	0.100	0.125	0.149	-0.027	0.052	0.132	0.168	0.173	0.177
Hangzhou	0.078	0.088	0.099	0.564	0.581	0.598	0.158	0.163	0.169
Ningbo	0.129	0.150	0.170	-0.022	0.050	0.123	0.148	0.150	0.153
Wenzhou	0.064	0.078	0.091	0.188	0.291	0.394	0.120	0.125	0.129
Taizhou	0.051	0.059	0.066	0.247	0.327	0.406	0.139	0.147	0.154
Hefei	0.079	0.098	0.118	0.009	0.074	0.139	0.142	0.146	0.151
Bengbu	0.173	0.236	0.299	-0.007	0.002	0.012	0.172	0.175	0.179
Anqing	0.090	0.116	0.143	-0.025	0.058	0.140	0.166	0.172	0.178
Fuyang	0.086	0.117	0.148	-0.020	0.014	0.047	0.132	0.136	0.141
Liuan	0.151	0.191	0.232	-0.001	0.001	0.003	0.169	0.172	0.174
Bozhou	0.140	0.181	0.222	-0.049	0.048	0.144	0.171	0.175	0.179
Fuzhou	0.063	0.086	0.109	-0.039	0.046	0.131	0.158	0.167	0.176
Putian	0.037	0.048	0.058	-0.001	0.000	0.002	0.154	0.167	0.179
Nanchang	0.076	0.093	0.110	0.055	0.164	0.273	0.145	0.150	0.155
Jiujiang	0.055	0.069	0.083	0.003	0.075	0.147	0.140	0.147	0.154
Xinyu	0.101	0.130	0.158	-0.015	0.125	0.265	0.157	0.162	0.166
Ganzhou	0.096	0.133	0.169	-0.010	0.015	0.041	0.147	0.152	0.157
Yichun	0.101	0.182	0.264	-0.005	0.112	0.228	0.154	0.160	0.166
Fuzhou	0.139	0.170	0.202	-0.006	0.041	0.089	0.177	0.180	0.183
Shangrao	0.160	0.201	0.242	-0.035	0.072	0.179	0.158	0.161	0.164
Zhengzhou	0.052	0.066	0.081	-0.009	0.001	0.011	0.144	0.152	0.159
Nanyang	0.094	0.138	0.181	-0.036	0.039	0.113	0.114	0.120	0.125
Shangqiu	0.139	0.187	0.236	-0.005	0.001	0.006	0.137	0.142	0.147
Xinyang	0.083	0.109	0.135	-0.048	0.179	0.405	0.145	0.150	0.156
Zhoukou	0.050	0.064	0.077	-0.008	0.020	0.047	0.154	0.164	0.175
Zhumadian	0.120	0.147	0.173	-0.005	0.001	0.007	0.160	0.164	0.168
Huangshi	0.097	0.109	0.121	0.008	0.070	0.131	0.086	0.088	0.090
Shiyan	0.049	0.060	0.072	0.495	0.544	0.592	0.117	0.123	0.128
Yichang	0.064	0.075	0.086	0.020	0.097	0.174	0.085	0.088	0.090
Xiangfan	0.059	0.070	0.081	0.138	0.256	0.375	0.122	0.127	0.132
Ezhou	0.059	0.070	0.082	0.037	0.048	0.060	0.111	0.114	0.118
Jingmen	0.060	0.073	0.086	0.462	0.490	0.519	0.116	0.120	0.124
Xiaogan	0.074	0.089	0.104	-0.011	0.134	0.279	0.085	0.087	0.089
Jingzhou	0.075	0.088	0.101	0.062	0.146	0.230	0.100	0.102	0.105
Huanggang	0.077	0.090	0.102	0.064	0.114	0.164	0.093	0.095	0.097
Xianning	0.137	0.268	0.399	-0.026	0.033	0.092	0.075	0.080	0.086
Suizhou	0.059	0.069	0.078	0.064	0.125	0.186	0.106	0.109	0.113
Changsha	0.055	0.063	0.070	0.029	0.035	0.041	0.140	0.146	0.151
Zhuzhou	0.086	0.106	0.126	0.029	0.071	0.112	0.155	0.160	0.164
Shaoyang	0.113	0.135	0.157	-0.017	0.121	0.260	0.168	0.173	0.178
Yueyang	0.074	0.087	0.100	0.012	0.036	0.060	0.124	0.128	0.132
Changde	0.066	0.088	0.110	-0.012	0.044	0.100	0.138	0.146	0.153
Loudi	0.102	0.117	0.132	-0.012	0.051	0.113	0.138	0.141	0.144
Zhuhai	0.084	0.103	0.123	-0.003	0.001	0.004	0.139	0.142	0.146
Huizhou	0.148	0.214	0.280	-0.004	0.001	0.005	0.164	0.168	0.172
Dongguan	0.077	0.108	0.138	-0.048	0.080	0.209	0.138	0.143	0.149
Chongqing	0.062	0.074	0.085	0.032	0.064	0.097	0.106	0.110	0.113
Xian	0.103	0.120	0.138	-0.032	0.090	0.212	0.134	0.137	0.140

Table A.5: γ computed using 3 methods for chosen 53 cities. Confidence intervals are 95%.

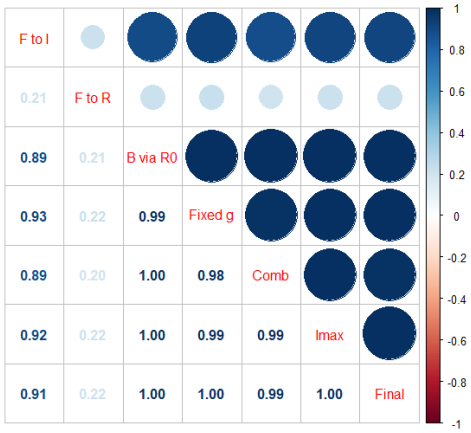
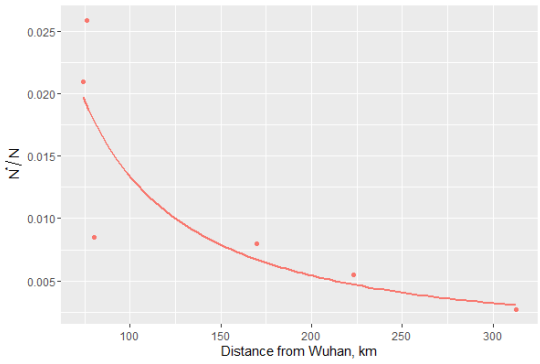
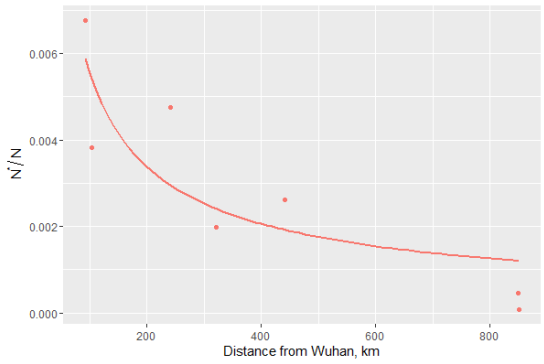


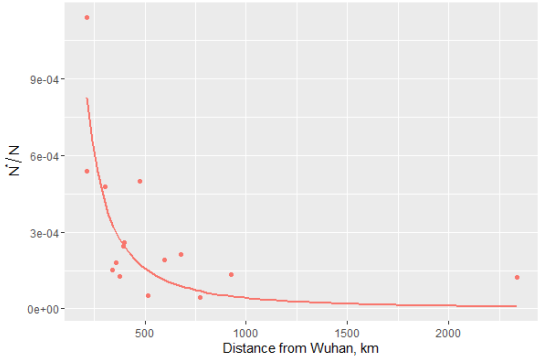
Figure A.9: A correlation matrix between optimum N^* results from each of the methods applied to Chinese cities COVID-19 data.



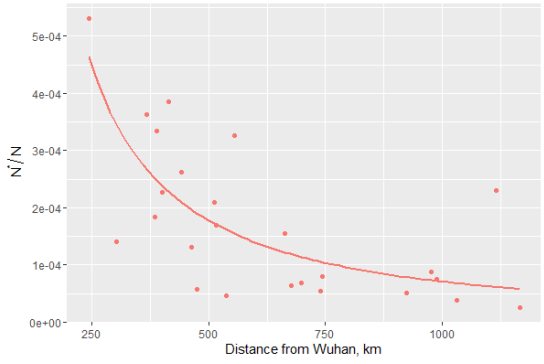
(a) Large outbreak locations $I_{max} \in [1000, \infty)$



(b) Big outbreak locations $I_{max} \in [250, 999]$



(c) Medium outbreak locations $I_{max} \in [100, 249]$



(d) Small outbreak locations $I_{max} \in [50, 99]$

Figure A.10: The relationship between distance between a city and Wuhan and N^*/N. N^* computed using fixed γ method. A curve correspond to non-linear fit of a function ax^b to data.