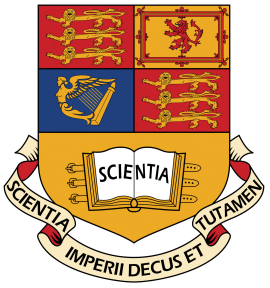
Imperial College London

Department of Computing

On-the-fly Modelling and Prediction of Epidemic Phenomena

Author:
Ionela Roxana Dănilă

Supervisor:
Dr. William Knottenbelt
Co-supervisor:
PhD student Marily Nika
Second marker:
Dr. Jeremy BRADLEY



A thesis submitted in fulfilment of the requirements for the degree of Master of Engineering in Computing

June 2014

"I simply wish that, in a matter which so closely concerns the wellbeing of the human race, no decision shall be made without all the knowledge which a little analysis and calculation can provide."

Daniel Bernoulli, 1760

Abstract

The modern world features a plethora of social, technological and biological epidemic phenomena. These epidemics now spread at unprecedented rates thanks to advances in industrialisation, transport and telecommunications. Effective real-time decision making and management of modern epidemic outbreaks based on model analysis depends on two factors: the ability to estimate epidemic parameters as the epidemic unfolds, and the ability to characterise rigorously the uncertainties surrounding these parameters. In this context, uncertainty should be understood as a statement about how well something is known, rather than being regarded as the act of not knowing.

The main contribution of this project is a generic Maximum Likelihood based approach towards on-the-fly epidemic fitting of SIR models from a single trace, which yields confidence intervals on parameter values. In contrast to traditional biological modelling techniques, our approach is fully automated and the parameters to be estimated include the initial number of susceptible and infected individuals in the population. Visualising the fitted parameters gives rise to an isosurface plot of the feasible parameter ranges corresponding to each confidence level.

We validated our methodology on both synthetic datasets generated using stochastic simulation, and real Influenza data. Fitting parameters to those trajectories revealed remarkable results. The model proved highly accurate in predicting from partial information on a single trace not only the time of the peak, but also its magnitude, and the tail of the infection. However, the "true" parameters were contained in the corresponding confidence bounds only for a relatively low proportion of the time, emphasising (a) the difficulty of obtaining accurate parameter estimations from a single epidemic trace and (b) the large potential impact of small random variations, especially those occurring early on in a trace.

Acknowledgements

First and foremost, I would like to thank to Dr. William Knottenbelt, his PhD student Marily Nika, and Dr. Jeremy Bradley for the input and guidance they have given me through the course of the project. Their dedication and passion for the subject were truly inspiring and contagious.

Secondly, I would also like to thank to my personal tutor, Prof. John Darlington, and to Mrs. Margaret Cunningham for their pastoral care and encouragement provided during my studies.

Last but not least, I would like to thank my parents for their continuous care and support throughout the years. Without them I would have not been able to pursue my dreams and to become the person I am today.

Contents

Chapter 1

Introduction

> As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.
>
> Albert Einstein

This project investigates uncertainty in epidemic modelling and presents a generic, fully automated method for on-the-fly epidemic fitting of SIR models from a single trace. It yields confidence intervals on parameter values that rigorously characterise the uncertainty inherent in their estimates. The modern era features a plethora of social, technological and biological epidemic phenomena. They spread at unprecedented rates due to advances in technology, transport and telecommunications. Mathematical modelling plays a key role in effective real-time decision making and management of modern epidemic outbreaks.

The ability to characterise uncertainty is absolutely critical in the context of policy and decision making. There is a popular misconception around the meaning of the term. Uncertainty is often regarded as not knowing. However, when it comes to decision making, it should be understood as a statement about how well something is known. As a general rule, uncertainty is inherent in science. Thus, to ignore or to minimise acknowledging its existence practically means to ignore science. Usually, there is a temptation to either focus only on the best estimates and ignore the less likely results, or to only consider the highly unlikely results based on extremely cautious assumptions. Both of these two approaches may lead to poor decision making. Instead, we should attempt to describe how far from the truth any given estimate is likely to be. Moreover, interpreting and framing of uncertainty may be subject to people's biases. It is therefore extremely important to develop a rigorous, scientific method that characterises uncertainty.

1.1 Motivation

Movement of disease constituted a major force in shaping the human history, with wars and migrations carrying infections to susceptible populations. Before World War II, more victims died due to microbes introduced by the enemy, than of battle wounds [1]. This was still a period of relative isolation across different communities. More recent times have allowed extensive contact between people around the world. Modern transport networks continuously expand in reach, speed of travel and volume of passengers carried, causing epidemics to spread further and faster than ever before. In the 14^{th} century, the Black Death travelled between 1 and 5 miles a day on average [2]. On the other hand, the severe acute respiratory syndrome outbreak of 2003 transmitted from Hong Kong to Hanoi, Singapore and Toronto within just a few days after the first infected case [3].

Gladwell [4] states that ideas, products, messages and behaviours spread in a similar manner to viruses, leading to social and technological epidemics. These phenomena are even more invasive due to the extensive coverage of internet and social media. Even so, they are based on the same three principles that explain how measles spread or why flu outbreaks occur every winter. Firstly, they have a contagious nature. Secondly, they may be triggered by seemingly inconsequential causes. Lastly and most important, there is one dramatic moment, the tipping point, when they begin to spread.

Our understanding of infectious disease dynamics has greatly improved in recent years thanks to mathematical modelling. Insights from this increasingly-important field enable policy-makers at the highest levels to interpret and evaluate data, in order to comprehend and predict transmission patterns. Compartmental models are widely used in epidemiology, allowing us to target control measures and use limited resources more efficiently. They reduce the population diversity to a few key characteristics, relevant to the phenomenon studied. For example, one of the most widely-known such models is SIR, which divides the population in susceptible, infected and recovered individuals. Parameters such as the rate of infection and the rate of recovery determine the behaviour of the model, but cannot be measured directly, hence they must be estimated in some way. Ultimately, the quality of a model is highly dependent on both the data used for parameterisation and the uncertainty present in the model outcomes.

One source of unreliability may arise when the data sets used for analysis are not entirely relevant to the hypothesis to test. A recent study published by two PhD students at Princeton University [5] states that Facebook will lose 80% of its users by 2017. One of the critical errors made in this non-peer-reviewed paper comes from applying a "correlation equals causation" principle. They deduced that a decline in the volume of Google searches for "Facebook" causes an ongoing decline in Facebook usage. However,

this decline does not prove anything considering that over half of Facebook's traffic comes from their mobile application at present. Indeed, since 2012, the number of active users kept growing, reaching today almost 1.2 billions [6]. Another source of error in their results is considering the Facebook phenomenon as a single outbreak, that starts by exponentially "infecting" people who then ultimately recover, causing the extinction of the epidemic. The user engagement strategy may be seen as a virus, but its mutations must not be omitted. In order to keep the engagement rates high, the company will continuously find new ways of attracting more users, generating each time new social and technological outbreaks.

Additionally, models are often developed and presented with insufficient attention to the uncertainties that underlie them. The authors of a recent study [7] analysed scientific papers, interviews, policies, reports and outcomes of previous infectious disease outbreaks in the United Kingdom. An extract from one of the scientific papers related to the dynamics of the 2001 UK foot and mouth epidemic is reproduced below:

"Relative infectivity and susceptibility of sheep and cattle. Experimental results agree with the pattern of species differences used within the model. Quantitative changes to the species parameters will modify the predicted spatio-temporal distribution of outbreaks; our parameters have been chosen to give the best match to the location of high risk areas. However this choice of parameters is contingent on the accuracy of the census distribution of animals on farms."

The purpose of their research was to ascertain the role uncertainties played in previous models, and how these were understood by both the designers and the users of the model. They found that many models provided only cursory reference to the uncertainties inherent in the parameters used. The study concludes that greater consideration of the limitations and uncertainties in infectious disease modelling would improve its usefulness and value.

Models provide epidemiologists with an environment able to record every detail of the disease spread, such that each individual component can be analysed in isolation to the whole system. However, every model has its limitations. There will always be an unknown or unknowable element in the system. For example, if we try to model Influenza, we need to account for factors such as movement and interaction of individuals, variability in susceptibility due to past infections, variations in transmission patterns caused by temperature and many more. We cannot capture all the different scenarios in order to predict the precise evolution of the epidemic. Instead, we should aim for providing confidence intervals on the parameters that determine the behaviour of the epidemic.

1.2 Objectives

The project aims to undertake on-the-fly parameter fitting as an epidemic unfolds, given regular observations in time of the number of infected individuals, and characterise the uncertainty inherent in the parameter estimates.

We will first consider least-squares-based techniques for parameter fitting to predict the future evolution of the epidemic, and answer questions such as "when will it peak?", "when will it have died out?", "how many people will be infected at a particular point in time", or "how many people need to be vaccinated to prevent an epidemic?"

Further, we aim to develop a rigorous maximum-likelihood-based methodology of characterising uncertainty. We consider uncertainty that comes from two sources: the stochastic evolution of the epidemic, and the parameters values, which are often unknown or imprecise. Traditional approaches used in biological epidemics require laborious manual work for index case identification, laboratory testing, contact tracing and report aggregation. The project will investigate to what extent a fully automated method could be deployed and, if possible, implement it.

We consider the challenges of estimating the initial number of susceptible and infected individuals in the target population, when these values are unknown. Currently, there is no principled way of doing this, as traditionally they are either supposed to be known, or can be estimated from the context [8]. However, in an era of social and technological epidemics, we argue that time and speed of movement make it infeasible to provide accurate estimates.

1.3 Contributions

This project made the following contributions:

- Investigation of on-the-fly parameters fitting as an epidemic unfolds, from a single trace using compartmental models.
- Implementation of a least-squares-based methodology for data fitting on SIR model, as en epidemic unfolds over time, tackling the challenge of unknown initial number of susceptibles.
- Implementation of a novel, fully automated maximum-likelihood-based methodology for data fitting on SIR model, as en epidemic unfolds over time, that provides

rigorous characterisation of uncertainty inherent in parameter estimates. We consider the challenges of applying this procedure when the initial number of susceptibles and infected individuals is unknown.

- A three-dimensional visualisation of the confidence intervals characterising parameters uncertainty in the SIR model when the number of susceptibles is unknown.
- Validation of the methodologies on both synthetic and real data.
- A paper co-authored with Thomas Wilding, Marily Nika and Dr. William Knottenbelt, and submitted to EPEW 2014, the 11$^{\text{th}}$ European Workshop on Performance Engineering, taking place in Florence, Italy, between 11-12 September 2014.

1.4 Report outline

Chapter 2 presents background information regarding infectious disease modelling. First, it introduces the laborious manual techniques traditionally used in developing disease control strategies. Subsequently, it highlights the role of mathematical modelling in epidemiology and describes in detail the main compartmental models widely used in this field. Further, it presents the main sources of uncertainty these models must account for. Next, an overview of other areas where compartmental models can be applied is given. Finally, we outline the main design decisions made regarding the programming language and additional libraries used to run the experiments.

Chapter 3 describes a least-squares based methodology for on-the-fly epidemic fitting on the SIR model, from a single trace. It provides mathematical details concerning the objective function, the optimisation technique, and the assessment of goodness of fit.

Chapter 4 introduces a novel, generic, and fully automated maximum-likelihood-based methodology for on-the-fly epidemic fitting on SIR models, from a single trace, that yields confidence intervals on parameter values. It represents a rigorous characterisation of the uncertainty inherent in parameter estimates. We first give mathematical details of the objective function, the optimisation technique, and the computation of confidence intervals. Then, we describe how we applied the methodology step-by-step for various vectors of unknown parameters. Finally, a three-dimensional visualisation of the confidence intervals is given, when dealing with three unknown parameters.

Chapter 5 presents the results of validating our methodologies on both synthetic and real data and a detailed discussion of their interpretation.

Chapter 6 concludes with a summary of the achievements and a discussion on future work.

Chapter 2

Background

2.1 Control of Epidemics

Improving control strategies and eradicating the disease from a population are the primary reasons behind studying infectious diseases. The Oxford English Dictionary defines an epidemic as "a widespread occurrence of an infectious disease in a community at a particular time." It can be described as a sudden outbreak of a disease, infecting a significant percentage of a population, that eventually disappears, usually leaving some of the individuals untouched. Management of epidemics involves a series of activities, from forecasting to investigation, control and prevention of future occurrences.

Traditional methods for disease control are applied after the extinction of an epidemic in order to better understand its dynamics from empirical data. These techniques have the potential of being highly efficient when dealing with a small number of cases. However, they become very tedious at a higher scale due to the laborious manual work usually involved, as discussed in Section 2.1.1.

During the course of an epidemic, it is extremely important to be able to predict the future course of the outbreak in real-time. In this context, prediction should be understood as both a quantitative approach, and an attempt of inferring what would happen under certain assumptions. Forecasting may not lead to a complete prevention of the epidemic, but it can control its severity and spread. Mathematical modelling plays a major role in accomplishing this, as discussed in section 2.1.2.

2.1.1 Traditional Methods

Traditionally, disease control strategies are developed after a series of laborious manual efforts. Epidemiologists collect data on symptoms, past medical history, laboratory testing, exam findings, and recent treatments that an infected individual have received, and also trace contacts between individuals. The aim is to identify the index case and the transmission network of the infection in order to understand its dynamics and make informed decisions for future prevention.

In epidemiology, an index case, also referred as patient zero, is considered to be the first documented case of a disease. Identifying these cases as soon as possible can provide significant information about the origin of the outbreak. It is also important to trace the pathways of the disease and construct its transmission network, which highlights details on how the disease spread. Infection tracing is an integral component of post-epidemic disease control policies. It aims to determine the source of infection for each case. The basic idea is to link each infected individual to both the one whom it caught the disease from, and the ones to whom they transmitted it to. In this way, the transmission network can be built. We discuss below the main traditional methods used to collect the required information.

Contact Tracing

Contact tracing is the process of identifying individuals who came in contact with an infected person. It aims to determine all potential transmission contacts from the index case. This methodology has many limitation. Firstly, it is highly laborious intensive and time consuming. Additionally, it fully relies on individuals being able to recall and provide complete, accurate information regarding their personal relationships.

Diary-based Studies

In contrast to contact tracing, diary-based studies attempt to record individual contacts as they occur. The advantage of this strategy is that the workload is shifted from researchers to the subjects, allowing a larger number of individuals to be tracked [9]. However, it also has a series of disadvantages. Firstly, the data collected is still at the discretion of individuals, hence its accuracy and consistency may vary. Secondly, it can be difficult for the coordinating researcher to organise all the information, as the identifiers of the contacts recorded may not be consistent.

2.1.2 Mathematical Modelling

Models represent a powerful tool in improving control strategies. They allow us to predict things such as the population-level epidemic dynamics from an individual-level knowledge of epidemiological factors, the long-term behaviour from the early invasion dynamics, or the impact of the vaccination on the spread of infection [10]. They provide a framework that conceptually explains how a system behaves. The rigour of the mathematical language used to define them can be combined with the simulation power of computers, providing means of studying of a system dynamics at a larger scale. There is an increasing interest in mathematical modelling in the epidemiological literature, as illustrated in Figure 2.1.

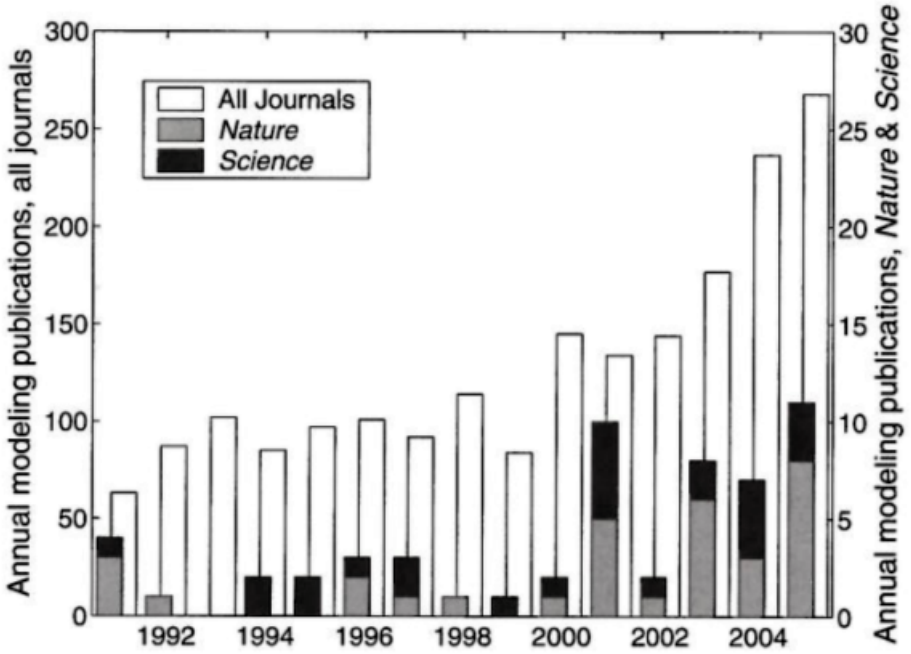


FIGURE 2.1: The importance and use of mathematical models in the epidemiological literature. Reproduced from [10].

Choosing the most appropriate model for a particular problem depends on various factors, such as the degree of precision required, the available data, or how fast the results are needed. The notion of *wrong, but useful*, attributed to statistician George Box, applies to all models in the sense that they require a set of simplifying assumptions. While the focus is on developing models that capture the essential features of a system, the usefulness of a model remains a subjective measure. The authors of [10] argue that formulating a model for a specific problem is a trade-off between three elements: accuracy, transparency, and flexibility. They define *accuracy* as the ability to reproduce the observed data and reliably predict future events. Whether a qualitative or a quantitative fit is necessary to measure accuracy fully depends on what purpose the model serves.

Gaining insight into the dynamics of the disease would require a qualitative fit, while establishing control policies would rather use a quantitative approach. The accuracy of a model is limited by computational feasibility, the modeller's understanding of the system in question, and the knowledge of the necessary parameters. *Transparency* is regarded as the ability to understand how the individual components of the system interact and influence the dynamics of the whole. The level of transparency usually decreases with the number of model components, as it becomes increasingly difficult to account for the role of each individual component. Finally, they define *flexibility* as a measure of how easily the model can be adapted to new situations. This proves to be essential when modelling diseases in an ever-changing environment.

According to [11], models can play three major roles in informing policy: prediction, extrapolation, and experimentation. Predictive models take a set of initial conditions and attempt to determine the future evolution of the epidemic, such as its size and location, in order to enforce appropriate control strategies. Models can also be used to construct the probable dynamics of a disease for a set of parameters by extrapolating from the known dynamics for another set of parameters. This can be useful when we are interested in studying the effects of relaxing or enhancing the control measures. Finally, models can be used to test various control strategies in a short period of time, by avoiding all the risks associated with testing during a real epidemic. One of the first times that models were used to support decision making during an epidemic was the 2001 foot-and-mouth disease outbreak in the UK. Three different models were used to investigate whether the epidemic was under control and assess to what extent targeted culling would be effective in reducing the spread of infection, in order to inform control measures.

No model is perfect and able to precisely predict the exact evolution of an epidemic. However, a good model is defined by two principles [10]. Firstly, it should be *suited to its purpose*. This means it should have a good balance of accuracy, transparency and flexibility. In other words, it must be as simple as possible, but no simpler, as it is often quoted in literature. Secondly, it should be *parametrizable from available data* for each of the features included. Hence, the definition of a good model is highly dependent on the context.

When developing a model, it is important to follow a series of steps in order to ensure that it is suitable for the problem it tries to address, and captures all the relevant information. Figure 2.2 illustrates the steps required in the development and use of a model.

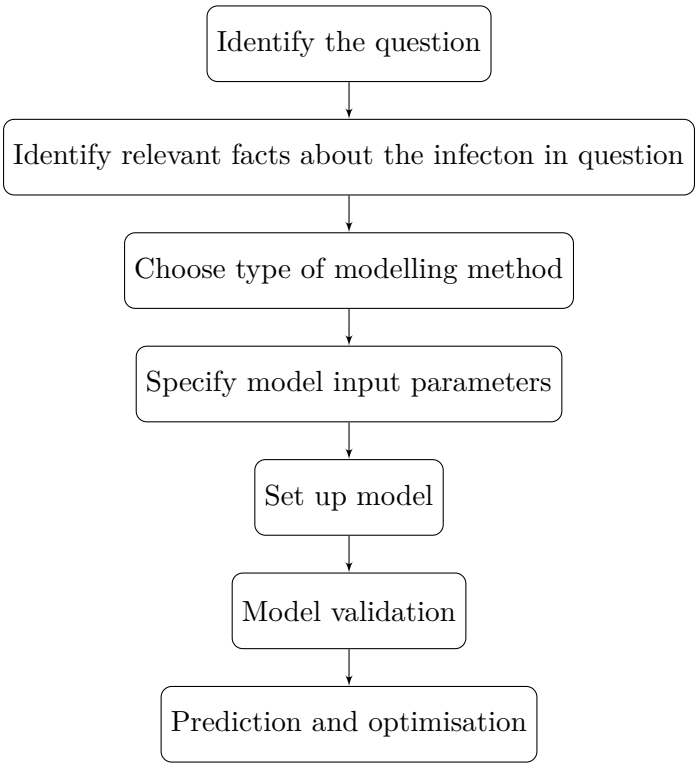


FIGURE 2.2: Steps in the development and use of a model. Adapted from [12].

Mathematical modelling has a long history in epidemiology. The first known result dates back from 1760 and is attributed to Daniel Bernoulli. It was an early attempt to statistically analyse the mortality caused by smallpox and defend the benefits of vaccination against it, a matter heavily debated at the time. In terms of modern mathematical epidemiology, the first contributions were made in the late 1880s, by Piotr Dimitrievich En'ko, a Russian physician whose probabilistic modelling and data analysis of measles epidemics anticipated the work of Reed and Frost in the 1920s.

One of the early triumphs in epidemiology is the approach based on simple compartmental models, developed between 1900 and 1935, having as contributors R.A. Ross, W.H. Hamer, A.G. McKendrick, W.O. Kermack and J. Brownlee. Compartmental models rely on two main assumptions. Firstly, it is assumed that the population under analysis can be divided into a set of compartments, depending on the stage of the disease development. Secondly, individuals are asserted to have equal probability to transit from one compartment to another. There are various questions that these models help us answer, including "how many individuals will be affected altogether and thus require treatment?", "what is the maximum number of people needing care at any particular time?" or "how long will the epidemic last?" [12].

2.2 Deterministic Compartmental Models

In a deterministic model, each state is uniquely determined by the parameters of the model, together with the previous state. Hence, for the same initial conditions, the model will behave exactly the same, such that each time we would observe an identical trajectory corresponding to the evolution of the epidemic.

2.2.1 SIR model

SIR is a compartmental model initially studied in depth by Kermack and McKendrick in 1927 [13]. It consists of dividing the population into three subpopulations: Susceptible, Infected and Recovered individuals, and uses Ordinary Differential Equations (ODEs) as a modelling formalism. It defines:

- $S(t)$ the number of individuals who are not yet infected at time t, but susceptible to become infected
- $I(t)$ the number of individuals who are infected at time t by contact with susceptibles at a rate β
- $R(t)$ the number of individuals who have recovered from the disease at time t at a constant rate γ

The model assumes that the size of each compartment is a differentiable function of time. It also considers a closed population, ignoring demographic processes such as births, deaths and migrations. There are two possible transitions taking place: S $\rightarrow$ I and I $\rightarrow$ R. The progression from S to I involves disease transmission at a rate βI, also known as the force of infection, where β is the probability of a contact between a susceptible and an infected individual resulting in infection. It ignores the intricacies related to the pattern of contact between individuals. The transition from I to R occurs at a recovery rate γ, assumed to be constant and equal to the inverse of the average infectious period.

Based on these assumptions, the flow diagram of the model is illustrated in Figure 2.3.

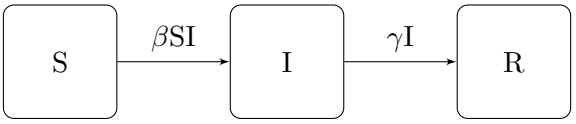


FIGURE 2.3: Flow digram for the SIR model. The boxes represent compartments of the population and the arrows indicate the flux between them.

The assumptions made above can be translated into an initial value problem, defined by the following set of differential equations:

$$\frac{dS}{dt} = -\beta SI \tag{2.1}$$

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

$$\frac{dR}{dt} = \gamma I \tag{2.3}$$

Differential equations used to model the transmission dynamics of a disease describe the events occurring continuously, opposite to difference equations that depict events taking place at discrete time intervals. Table 2.1 presents a comparison between difference equations, describing the number of susceptible, infected and recovered individuals at time t, and differential equations, illustrating the rate of change in the number of individuals in each compartment at time t.

Difference equations (number)	**Differential equations (rate)**
$S_{t+1} = S_t - \beta_t S_t I_t$	$\frac{dS}{dt} = -\beta(t)S(t)I(t)$
$I_{t+1} = I_t + \beta_t S_t I_t - \gamma_t I_t$	$\frac{dI}{dt} = \beta(t)S(t)I(t) - \gamma(t)I(t)$
$R_{t+1} = R_t + \gamma_t I_t$	$\frac{dR}{dt} = \gamma(t)I(t)$

TABLE 2.1: Comparison between difference and differential equations for the SIR model.

The use of differential equations avoids issues regarding the time step size that would arise in difference equations. As the size of the time step increases, the predicted epidemic curve becomes less and less smooth, and can even produce nonsense results. On the other hand, with the decrease of time step size, the model becomes closer to describing events that occur in continuous time, hence the predicted epidemic curve becomes smoother. This phenomenon is described in Figure 2.4.

The initial values of the SIR model must satisfy the following conditions:

$$S(0) = S_0 > 0 \tag{2.4}$$

$$I(0) = I_0 > 0 \tag{2.5}$$

$$R(0) = 0 \tag{2.6}$$

and at any time t, $S(t) + I(t) + R(t) = N$, where N is the total population size.

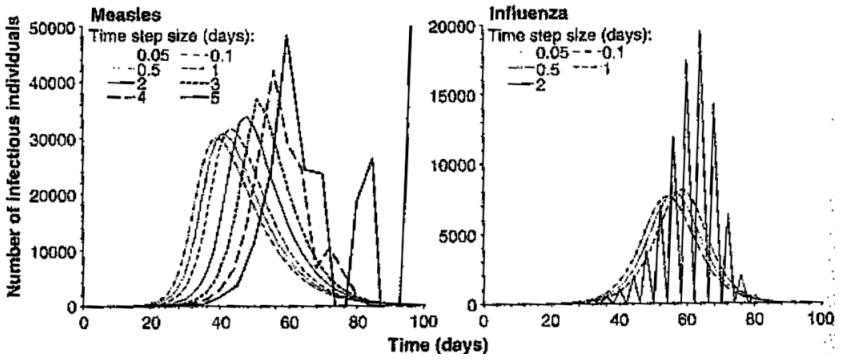


FIGURE 2.4: Comparison between predictions of the number of infectious individuals for measles and influenza, using time steps ranging between 0.05 and 5 days. Reproduced from [12].

An example of epidemic evolution, generated with parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S_0 = 600$, $I_0 = 60$ and $R_0 = 0$ over a period of 60 days are presented in Figure 2.5.

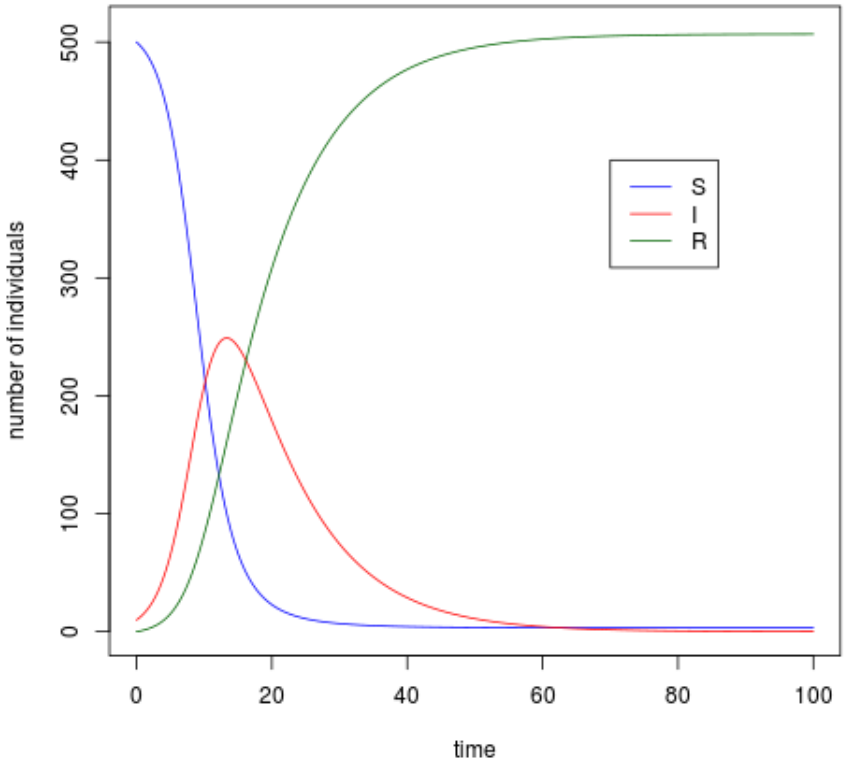


FIGURE 2.5: Sample run of the SIR model with parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S_0 = 500$ and $I_0 = 10$ over 100 days.

A detailed analysis in [14] shows that the model is mathematically and epidemiologically well posed, which means that there exists a unique solution and its behaviour changes continuously with the initial conditions. The trajectories of the model's dynamics in the phase plane are represented geometrically in Figure 2.6. Despite being

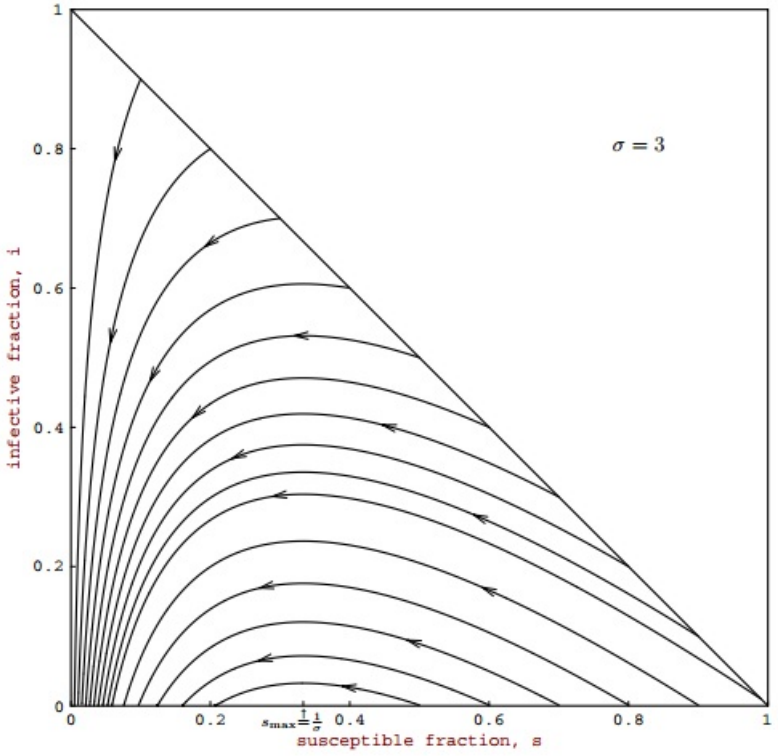


FIGURE 2.6: Phase plane portrait for the SIR model with contact rate $\sigma = 3$. Reproduced from [14].

extremely simple, we cannot solve this model analytically. However, it highlights important qualitative principles in epidemiology that help us learn about the behaviour of its solution. Firstly, it can be inferred whether an epidemic will occur or not based on the threshold phenomenon, notion closely related to the concept of basic reproductive ratio, described in detail in Section 2.2.3. Secondly, a rather counter-intuitive result regarding the long-term state of the infection can be derived, as described in Section 2.2.4.

2.2.2 Other Models

Many other models were developed from the classic SIR model, in order to allow various behaviours to be modelled. Some of them denote infections that are strongly immunizing, while others were developed for infections that do no give rise to immunity. These approaches ignore heterogeneities related susceptibility to infection, transmission through contact networks, or immunological responses.

SI model The SI model was developed to account for the case when the infection can induce mortality. S and I remain the subpopulations of susceptibles and infected, respectively. We also consider ρ as the probability of an infected individual to die before recovery or from natural causes, which takes values between 0 and 1. Additionally, μ is the rate of natural mortality. Mathematically, the model is described by the set of ODEs:

$$\frac{dS}{dt} = -\beta SI \tag{2.7}$$

$$\frac{dI}{dt} = \beta SI - \frac{(\gamma + \mu)}{1 - \rho} I \tag{2.8}$$

There are other variations to this model that take into account various stages at which an infection may produce mortality.

SIS model The previous described models illustrate the dynamics of epidemics that either confer immunity after recovery or induce death. The SIS model captures those epidemics that don't confer life-lasting immunity, such that an individual recovered from the infection becomes susceptible again. The long term persistence is guaranteed by the loss of immunity, which always replenished the susceptibles pool. The following pair of ODEs describe the model:

$$\frac{dS}{dt} = \gamma I - \beta IS \tag{2.9}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2.10}$$

The parameters remain similar to the ones in the previous section, except that $S+I = N$, where N is the total size of the population.

SEIR model The SEIR model introduces a new category of individuals E, namely Exposed, consisting of individuals who are infected, but not yet infectious. Taking the average duration of this latent period is $\frac{1}{\alpha}$, the model is given by the following differential equations:

$$\frac{dS}{dt} = -\beta SI \tag{2.11}$$

$$\frac{dE}{dt} = \beta SI - \alpha E \tag{2.12}$$

$$\frac{dI}{dt} = \alpha E - \gamma I \tag{2.13}$$

$$\frac{dR}{dt} = \gamma I \tag{2.14}$$

We also assume that $S + E + I + R = N$. Compared to the SIR model, it has a slower growth rate due to the fact that individuals must belong to the Exposed subpopulation before being able to transmit the infection.

MSEIR model A more general model is MSEIR, which also includes a category of individuals that are passively immune since their mothers developed some type of immunity. It is suitable for modelling a directly transmitted disease with permanent immunity after recovery, in a population with variable total size. It translates to the following system of differential equations:

$$\frac{dM}{dt} = b(N - S) - (\delta + d)M \tag{2.15}$$

$$\frac{dS}{dt} = bS + \delta M - \beta SI/N - dS \tag{2.16}$$

$$\frac{dE}{dt} = \beta SI/N - (\epsilon + d)E \tag{2.17}$$

$$\frac{dI}{dt} = \epsilon E - (\gamma + d)I \tag{2.18}$$

$$\frac{dR}{dt} = (b - d)N \tag{2.19}$$

where b and d are the constant rates of birth and death, respectively.

2.2.3 Basic Reproductive Ratio

For the SIR model, a famous result highlighted by Kermack and McKendrick [13] is known in the literature as the threshold phenomenon. It states that in order for an epidemic to spread, the initial number of susceptibles must exceed a certain threshold, equal to γ/β. The value of the threshold is derived by re-writing Equation 2.2 as:

$$\frac{dI}{dt} = I(\beta S - \gamma) \tag{2.20}$$

In the initial stage, after $I(0)$ infected individuals are introduced in the population, the infection becomes extinct if $dI/dt < 0$, which is equivalent to $S(0) < \gamma/\beta$. This threshold is referred to as the relative removal rate, which must be small enough in order to allow the infection to spread.

The inverse of this rate is called *basic reproductive ratio* R_0, and constitutes one of the most important measures in epidemiology. It is formally defined in [15] as:

the expected number of secondary infections arising from a single individual during his or her entire infectious period, in a population of susceptibles .

Figure 2.7 illustrates implications of a basic reproductive ratio $R_0 = 4$. At each consecutive time point, each individual can transmit the infection to up to four others.

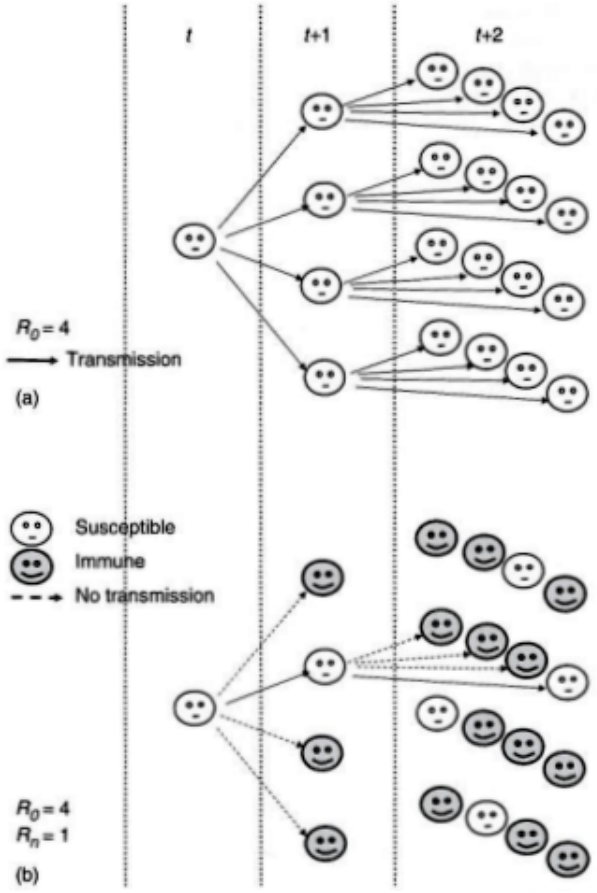


FIGURE 2.7: Implications of a basic reproductive ratio $R_0 = 4$.
Reproduced from [12].

In an entirely susceptible population (a), the incidence increases exponentially. In a population that is 75% immune (b), only 25% of the contacts lead to infection.

From the definition, it follows immediately that an epidemic will spread if and only if $R_0 > 1$, which is just another way of expressing the threshold phenomenon. In its simplest form, R_0 is mathematically expressed as:

$$R_0 = \frac{\beta}{\gamma} N = cpD \tag{2.21}$$

where
β = infection rate
γ = recovery rate
N = total size of the population
c = contact rate
p = transmission probability given contact
D = duration of infectiousness

However, estimating R_0 from individual parameters is not always feasible, as they might be unknown or impossible to estimate. Alternatively, the basic reproductive ratio can be estimated from epidemic time series data [16].

If the exponential growth rate of the initial phase r is available, then:

$$R_0 = 1 + rD \tag{2.22}$$

If the doubling time of the number of infected individuals t_d is known, then:

$$R_0 = 1 + \frac{Dln2}{t_d} \tag{2.23}$$

If we consider s_0 the number of susceptibles before the outbreak and s_α the number of susceptibles after the epidemic dies out, then:

$$R_0 = \frac{ln(S_0) - ln(s_\alpha)}{s_0 - s_\alpha} \tag{2.24}$$

Table 2.2 presents examples of various diseases and their corresponding estimated values for the basic reproductive ratio. Because R_0 depends on both the disease and the host populations, differences in demographics or contact rates may lead to different estimated values for the same disease.

Infectious disease	**Host**	**Estimated R_0**	**Reference**
Rabies	Dogs Kenya	1.1 - 1.5	Smith (2011)
Tuberculosis	Cattle	2.6	Goodchild and Clifton-Hadley (2001)
1918 Pandemic Influenza	Humans	2 - 3	Mills et al. (2004)
Foot-and-mouth Disease	Livestock farms UK	3.5 - 4.5	Ferguson et al. (2001)
Rubella	Humans UK	10 - 12	Anderson and May (1991)
Measels	Humans UK	16 - 18	Anderson and May (1982)

TABLE 2.2: Estimated basic reproductive ratios for various diseases. Adapted from [12].

2.2.4 Epidemic Burnout

Another important result derived from the SIR model is related to the long-term state of the epidemic. Firstly, it has been observed that there will always be a certain number of susceptible individuals who do not get infected. Mathematically, this can be derived

by dividing Equation 2.1 by Equation 2.3:

$$\frac{dS}{dR} = -\frac{\beta S}{\gamma} = -R_0 S \tag{2.25}$$

and integrating with respect to R:

$$S(t) = S(0)e^{-R(t)R_0} \tag{2.26}$$

This shows that S always stays positive. The conclusion that emerges from this result is rather counter-intuitive: *the chain of transmission eventually breaks due to the decline in infectives, not due to a complete lack of susceptibles* [10].

2.3 Uncertainty Sources

The application of compartmental models in epidemiological modelling is accompanied by concerns regarding the degree of uncertainty prevailing in their use. There are two main sources of uncertainty that we consider, discussed below.

2.3.1 Stochastic Uncertainty

Stochastic uncertainty arises from the randomness present in the evolution of an epidemic. If an infectious disease outbreak would re-occur, we would not observe the exact same number of infected individuals at the same time. This intuitively suggests that a stochastic model is always desirable, being more realistic. However, the magnitude of the fluctuations depend on the population size. A large population reduces the fluctuation level, hence a deterministic model can provide a good approximation. When addressing small populations or diseases with reduced level of incidence, stochasticity can make a tremendous difference. It introduces variances and co-variances that may lead to chance extinction of the disease.

Computationally, stochastic uncertainty can be simulated using Gillespie's discrete-event simulation algorithm (SSA) [17]. This is applicable to systems that can be modelled as a continuous-time Markov process whose probability distribution obeys a so called "master equation". It produces single realisations of the stochastic process that statistically agree with the master equation.

2.3.2 Parameter Uncertainty

Parameter uncertainty relates to the fact that the outcomes of fitting data against a model are themselves uncertain, because they are quantities estimated from subjective information. Factors such as the sample size informing that estimate, and variance in the data contribute to determining the level of parameter uncertainty.

2.4 Other Applications of Epidemiological Models

2.4.1 Social Network Analysis

Online Social Networks (OSN) represent web-based services that allow users to have a presence via their individual profile, build a list of connections and interact with them. The concept originated in the 1960s with Plato, a computer-based education tool developed at University of Illinois, but the viral growth and commercial interest present today only started after the advent of the Internet. Nowadays, online social networking is a mass adoption phenomenon. For example, public data on Facebook's website, the largest OSN at present, reveals 1.23 billion monthly active users, with an average of 757 million users that log in daily as of December, 31 2013. Every 20 minutes, 1 million links are being shared, 2 million friends requested and 3 million messages sent on average [6].

Social network analysis has a wide range of applications across multiple disciplines such as data aggregation and mining, network modelling, user attribute and behaviour analysis, location-based interaction, social sharing and filtering, recommendation systems development, or link prediction. In the private sector, businesses use OSN analysis for to fulfil their marketing and business intelligence needs, while in the public sector it serves to the development of leader engagement and community-based problem solving. Also, law enforcement and intelligence institutions make use of this technique in fighting and preventing crime.

The relationships among social entities and the patterns and implications they have on content spreading developed researchers' interest in OSN analysis. The online environment promotes viral dissemination of information, creating powerful electronic world-of-mouth effects that result in the birth of online trends [8].

2.4.2 Economic cycles

The idea of adopting in economics tools and techniques from biology is not new, being firstly highlighted by the neoclassical economist Alfred Marshall in the preface to his

Principles of Economics (1890): "the Mecca of the economist lies in economic biology". If we consider the economy as a heterogeneous system comprising of different typologies of agents that interact, influence each other and have different levels of knowledge about the environment and each other, then biology can provide the necessary tools to explain various behaviours of agents.

It is interested to observe how compartmental models could be used to model the behaviour of economical phenomena, such as *business cycles*. The standard definition of the term was given by Burns and Mitchell in 1946 [18]: business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle. The concept is illustrated in Figure 2.8.

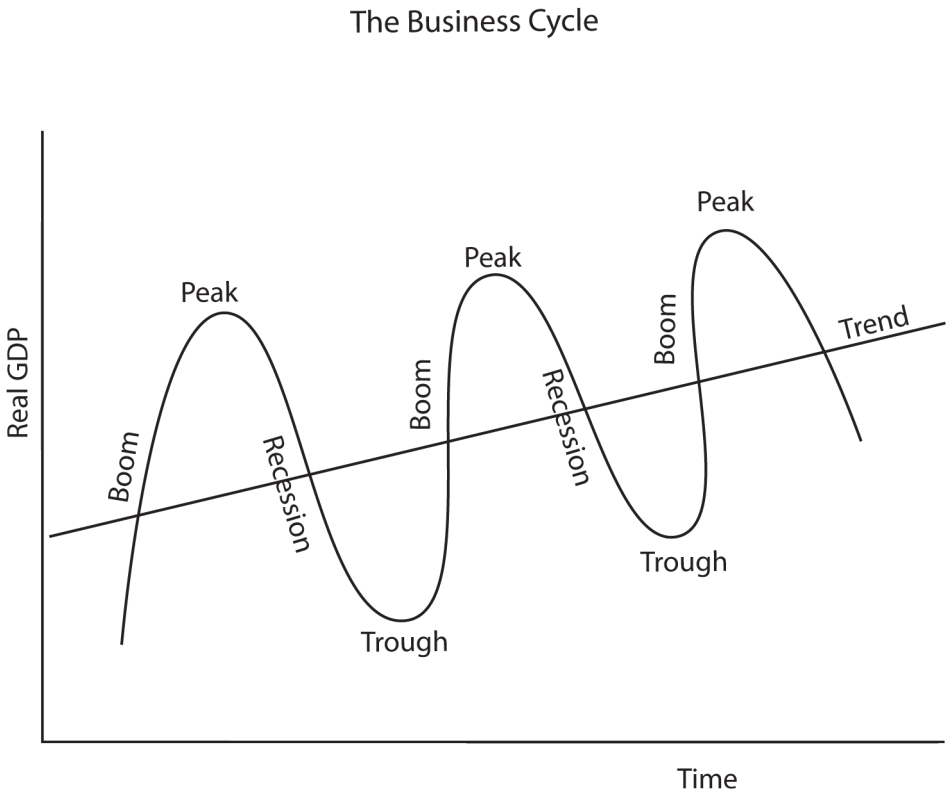


FIGURE 2.8: A basic illustration of the economic/business cycle

Strictly speaking, business cycles capture the upwards and downwards economical movements that occur around a long-term growth trend. Beside being called "cycles", these fluctuations often prove unpredictable and finding an explanation for them is one the primary concerns in macroeconomics.

In economic literature we cannot find many applications of compartmental models. One of the few existing works related to this subject belongs to the Nobel laureate Gary Becker, who divides the population into various subclasses and analyses the causes of deceleration of the flow from one category to another, without specifically using the word ”compartment”. Another approach, closer to the epidemiological models, is the one of Aoki [19], who introduces the notion of clusters, conceptually similar to compartments. A big difference, however, is the stochastic approach taken, which significantly increases the complexity of the mathematical tools required. A direct application of compartmental models in economy is studied in [20], by introducing a simple model describing how the government could operate in order to reach a certain goal as cost-efficiently as possible.

2.4.3 Retail Sales

Another less explored area where epidemic models could be applied is retail sales. The awareness of a new product spreads among customers similar to how a disease spreads from an individual to another. Epidemic models could take various factors into account, such as the impact of previous negative experiences with a product, and use the number of early adopters to project peak sales or sales volume levels over time.

This subject is little explored in the literature, although we can see how compartmental models could be constructed. For example, in the context of online retail sales, [21] suggests to assign all major products categories to one of three market share groups: high, medium or low penetration potential. This division should be based on the suitability of the product to the online medium and its historical online success to date. However, the authors of the report do not build a set of ODEs, but derive a specific logistic growth function for each category.

2.4.4 Computer Viruses

The Code Red worm incident that happened in July 2011 raised awareness regarding the urge to build models for analysing how Internet viruses propagate. Researchers in [22] developed a general Internet worm model based on the classic epidemic SIR model, taking into account two major factors. Firstly, they considered the dynamic countermeasures taken by users in removing susceptible and infectious computers. The second factor taken into account is the slowed down infection rate, as a consequence of the congestions to some routers caused by its large-scale propagation. The results of fitting the data for this model are illustrated in Figure 2.9.

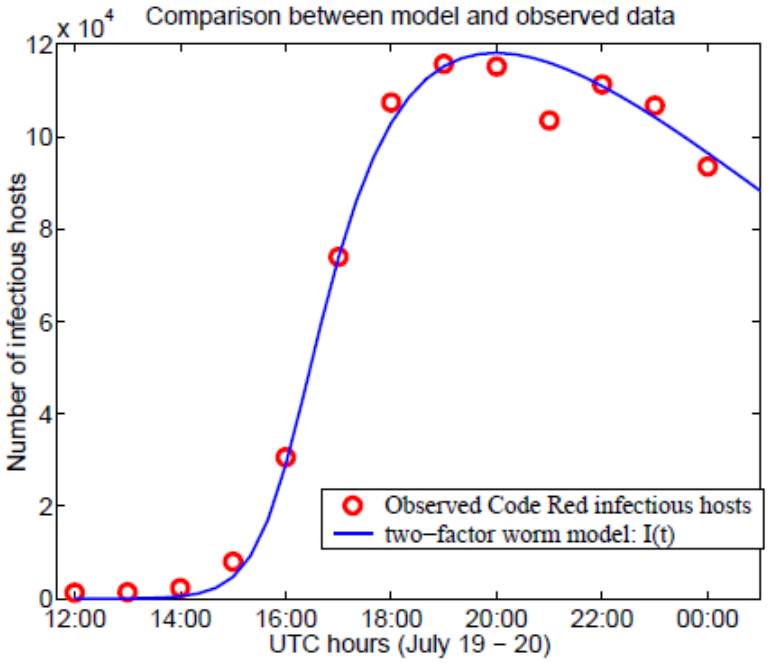


FIGURE 2.9: Comparison between observed data and the two factor model. Reproduced from [22].

More recent research revealed some issues present in previous epidemic models, highlighted in [23]:

- models including an exposed compartment do not take into account that an computer can infect other computers immediately after getting infected
- models including all infected computers into one compartment do not take into account the difference between latent and breaking-out computers
- models including a permanent immunity compartment do not take into account that previously infected computers are prone to infection by new versions of the virus
- most of the models do not take into account the effect of removable storage media
- all models assume that a computer is uninfected when connected to the Internet

Various models have been developed to tackle each of the five mentioned issues individually, until the authors of the same study proposed a novel epidemic model that accounts for all of them.

2.5 Development Environment

2.5.1 Programming Language

In terms of programming language, the obvious choices were Matlab and R, both being powerful and widely used for statistical modelling and data analysis. With no experience in either of the two languages, *R* was chosen after comparing their advantages and disadvantages.

The main difference between the two comes from the fact that Matlab is a commercial software, while R is open source. Therefore, R is free, allowing anyone to use it and contribute to its enhancement. Its vast user community of over 2 million people is constantly adding new packages, enriching its set of functionalities. At present, it is the most comprehensive statistical analysis tool available, making it ideally suited for the purpose of this project. Statisticians were the ones who created R. This constitutes a major advantage because it means data analysis lies at the very heart of the language. However, it is not as well documented as Matlab, and although many introductory tutorials are available, none of them are comprehensive enough. It is not straightforward to obtain a clear overview of the available functionalities, and looking for the right package can be time consuming.

R admittedly has a steep learning curve. Apart from the poor documentation, implementation details, such as silent coercion or sometimes misleading textual presentation of objects contribute to this phenomenon. Mistakes are very easily made and careful consideration must be given in order to avoid common pitfalls. Despite this aspect, R was still preferable over Matlab due to its data frames. A data frame is a core data structure, similar to a matrix in Matlab, with two primary advantages: firstly, the rows and columns can be named rather than being referred by index, and secondly, each column can hold a different data type.

Similar to Matlab, R is cross-platform compatible, being available under various operating systems and architectures. It also integrates well with many other tools. For example, it can import data from sources such as CSV, Microsoft Excel, MySQL. It can also produce graphics output in PDF, JPG, PNG, and SVG formats, and table output for LATEX and HTML. Compared to Matlab, that can produce high quality interactive plots, R's visualisation capabilities are better suited for exploratory analysis, which plays a major role in this project.

One of the main disadvantages of R concerns memory management. R holds objects in virtual memory, and limits are imposed on the amount of memory that can be used. The limitations apply to the size of the heap and the number of cons cells allowed.

The environment may further limit the user address space of a single process and the resources available to a single process. Because many R commands give little thought to memory management, it can quickly run out of resources. However, this usually happens when working with huge data sets, which is beyond the scope of this project.

In terms of performance, both R and Matlab are fast when it comes to mathematical operations on arrays, which are the main data structures used throughout the project. However, they have slow language interpreters, discouraging complex abstractions.

Another possible choice would have been Python, which is overall a better programming language than both R and Matlab. Its object oriented and functional nature, together with libraries such as Numpy, Scipy, statsmodels, and matlibplot make it a powerful statistical tool. However, it lacks a strong community of mathematicians, so many of the functionalities already existing in Matlab and R are not yet available.

2.5.2 Additional Libraries and Tools

R: The main R packages used in this project are: *deSolve*, *bbmle*, and *GillespieSSA*.

The package *deSolve* provides general solvers for initial value problems of first order Ordinary Differential Equations (ODEs) systems, assuming a full or banded Jacobian matrix. It also includes fixed and adaptive time-step Runge-Kutta solvers, as well as the Euler method.

The package *bbmle* provides tools for general maximum likelihood estimations. It extends the *stats4* default package, being superior to it in some respects. Firstly, the functions are more robust, with additional warnings that allow certain computations to return, rather than stop with an error. Secondly, it allows for more parameters to be passed to the negative log-likelihood function via a data argument. Additionally, for simple models an in-line formula may be passed to the optimisation procedure, instead of defining a separate negative log-likelihood function.

The package *GillespieSSA* provides an interface to various stochastic simulation algorithms for generating simulated trajectories of finite population continuous-time models. The interface is simple to use, intuitive and easily extensible. Currently, it implements various Monte Carlo procedures for Gillspie's Stochastic Simulation Algorithm (SSA), including both direct and approximate methods.

Matlab: Although the main implementation language was chosen to be R, it proved a challenging environment for neat 3D surface plots. Hence, we used Matlab to produce isosurface plots.

Chapter 3

Fitting Procedure using Least Squares

This chapter describes a least-squares-based epidemic fitting procedure of an SIR model, as the outbreak unfolds over time. We first introduce the basic idea behind this method. Then, we describe the model used for fitting, and provide mathematical details regarding the objective function and the optimisation technique. Finally, we outline a measure to assess the goodness of fit. Our methodology tackles the challenge of considering the initial number of susceptibles unknown.

Least Squares (LS) is a simple approach to investigate the evolution of epidemic dynamics over time and estimate the parameters values, first documented by Gauss around 1794. We assume that the only source of variability in the data comes from measurement errors and that its variance is constant, with a symmetrical distribution. Under these circumstances, Least Squares constitutes a statistically appropriate method for estimation, being a procedure that allows finding approximate solutions of overdetermined systems, i.e. systems that have more equations than unknowns. The basic idea behind it is to test different values of parameters in order to find the best fit model for the given data set. However, the robustness of least squares is highly dependent on how close to the model are the data points. Thus, outliers can cause inaccurate estimates.

3.1 Model

In order to fit our SIR model using Least Squares, we analyse epidemic curves from data reporting incidence of the disease through time. An example of such data is illustrated in Figure 3.1.

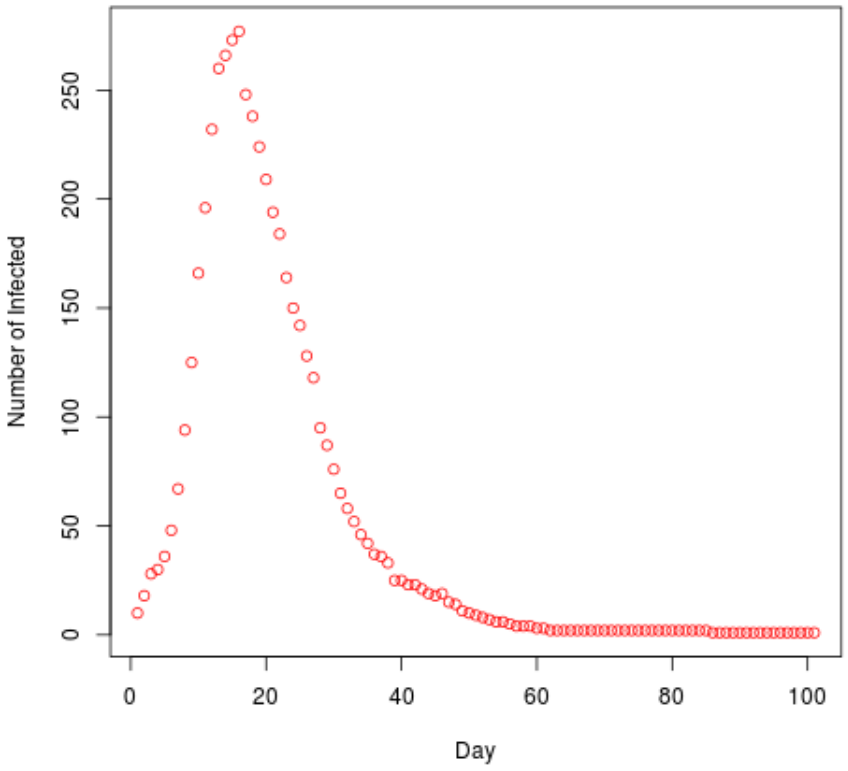


FIGURE 3.1: Daily number of infected individuals over a period of 100 days.

The equations describing the SIR model cannot be solved analytically, hence numerical integration methods are required. We solve the differential equations numerically using the function *ode* in the R solver package *deSolve*. The function requires as parameters a set of initial values, a time sequence for which output is wanted, and a model definition. A simplified R implementation of the SIR model is presented in Figure 3.2. For clarity, we omit here extra checks, which ensure that data has the right type and it lies within a sensible range of values.

```
sir.model <- function (t, x, params) {
  S <- x[1]
  I <- x[2]
  R <- x[3]

  beta <- params[1]
  gamma <- params[2]

  dS <- -beta*S*I
  dI <- beta*S*I-gamma*I
  dR <- gamma*I

  c(dS, dI,dR)
}
```

FIGURE 3.2: SIR model equations implemented in R.

As integration method we use the integrator *lsoda* provided in the same package. This solver is robust due to its automatic detection of stiffness, i.e. property that makes unstable certain numerical methods for solving equations, unless an extremely small step size is being used. Its implementation uses linear multi-step methods that approximate the derivative of a given function using information computed in previous steps. In particular, an explicit multi-step Adams method is applied for non-stiff systems, and the Backward Differentiation Formulas (BDF) method for the stiff ones. In terms of accuracy, the default relative tolerance and absolute tolerance are equal to 10^{-6}, determining the error control performed by the solver. Alternatively, a maximum value for the integration step-size may be specified.

Figure 3.3 illustrates the SIR model trajectory for parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S0 = 500$, $I0 = 10$, $R0 = 0$, during a period of 100 days in a closed population.

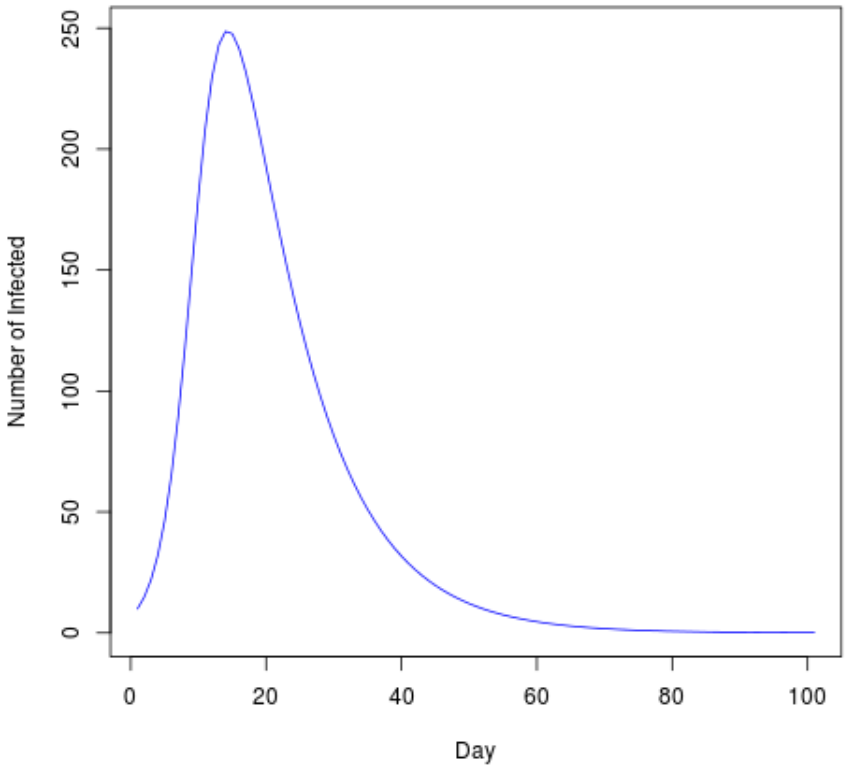


FIGURE 3.3: Trajectory prediction for SIR model with parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S0 = 500$, $I0 = 10$, $R0 = 0$

3.2 Objective Function

The first step in trajectory matching is defining an objective function. Least Squares finds a solution by minimising the sum of the squares of the errors. This is also one of its *limitations*. Using *the squares* of the error differences in the presence of outlying points may lead to a disproportionate effect in the fit, property which is usually not desirable. Outliers can potentially cause the estimates to be outside a desired range of accuracy. The method is therefore only as robust as the observed data points are close to the model.

The *basic idea* is to find estimates of the parameters that minimise the squared offsets of the model predictions from the observed data. Algebraically, this is equivalent to minimising:

$$S = \sum (y_i - f(x_i, \theta))^2 \tag{3.1}$$

where y_i is the observed value, and $f(x_i, \theta)$ is the model function, with θ being the vector of unknown parameters.

Figure 3.4 illustrates the R implementation for our objective function, that computes the squared differences (sum of squared errors) between the observations and any parameterisation of the model.

```
sir.sse <- function (params, data) {
  logBeta  <- params[1]
  logGamma <- params[2]

  pred <- as.data.frame(ode(y=c(S=500, I=10, R=0),
                            times=data$Time,
                            sir.model,
                            parms=exp(c(logBeta,logGamma))))
  obs <- data[,2]
  sse <- sum((pred$I-obs)^2)
}
```

FIGURE 3.4: R implementation of the objective function that computes the sum of least squared errors.

It is important to highlight the *modelling trick* used in the implementation of the objective function. We know that β and γ must always be positive, as they represent the rate of infection and the rate of recovery, respectively. Originally, our optimisation procedure would have searched over the entire range between $-\infty$ to $+\infty$. Intuitively, we want to constrain the search space to one that is meaningful in the context of our model, namely 0 to $+\infty$. We achieve this by parameterising the objective function in terms of $\log(\beta)$ and $\log(\gamma)$. Alternative approaches based on more sophisticated constraining algorithms are available, but they may lead to problems such as stability at the boundaries.

3.3 Optimisation Technique

The second step computes the parameter estimates that minimise the objective function. To achieve this, we use the function *optim* in the R package *stats*, which provides robust algorithms for general-purpose optimisations. The technique we selected is based on the *Nelder-Mead algorithm*, a widely used method in multidimensional unconstrained optimisation. It falls under the general class of direct search methods, as it does not involve any explicit or implicit derivative information. This makes it suitable to solve optimisation problems even when the objective function is not smooth [24].

Figure 3.5 illustrates an example of curve fitting using Least Squares. The observed data was generated synthetically with parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S0 = 500$, $I0 = 10$, $R0 = 0$ over a period of 100 days within a closed population.

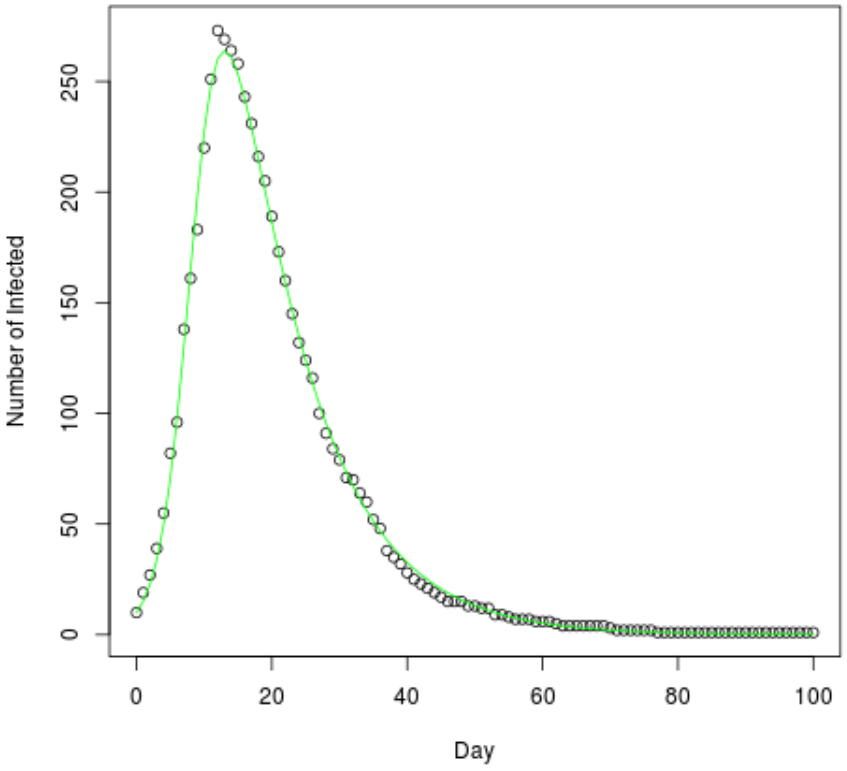


FIGURE 3.5: Curve fitting using Least Squares.
Original parameter values: $\beta = 0.001$, $\gamma = 0.1$.
Estimated parameter values: $\beta = 0.001025721$, $\gamma = 0.093358939$.

3.4 Goodness of Fit

Finally, after fitting the data with the model, we evaluate the goodness of fit. We aim to assess how well a chosen set of parameters fits the observed data by identifying the discrepancies between them. After a visual examination, we make use of the *coefficient of determination*, denoted R^2.

R^2 is a statistical measure, usually reported in the context of regression. It determines how much of the total variation present in the observed data is explained by the model. The sample variance is proportional to the total sum of squares SS_{tot}, given by Equation 3.2. To measure how far from the observed data are the estimates, we compute the residual sum of squares SS_{res}, using Equation 3.3.

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{3.2}$$

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{3.3}$$

where f_i are the model predictions, y_i are the observed data points and $\bar{y}$ is the mean of the observed data, given by Equation 3.4.

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3.4}$$

Based on these measures, the coefficient of determination is defined by Equation 3.5:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.5}$$

Generally, R^2 ranges between 0 and 1. Its *interpretation* denotes the degree of improvement the model has made over the average of the observed data. Hence, the closer R^2 is to 0, the least agreement between the actual and estimated values is observed. The closer R^2 is to 1, the better explained is the variability in the data. However, from the definition we notice that R^2 can take negative values if $SS_{res} > SS_{tot}$. In this situation, it can be inferred that the mean of the observed data provides better estimates than the ones of the fitted model. A key *limitation* of R^2 is that it cannot determine whether the model prediction and estimates are biased. This is why we must also examine the residual plots.

Chapter 4

Uncertainty Characterisation using Maximum Likelihood

This chapter describes a new maximum-likelihood-based epidemic fitting procedure of SIR model as the outbreak unfolds over time, yielding confidence intervals on the estimated parameters. It is a generic, fully-automated methodology for rigorous characterisation of the uncertainty inherent in the estimated values. We first introduce the basic idea behind Maximum Likelihood, providing mathematical details about the objective function and the optimisation technique used. Then, we discuss how confidence intervals characterise uncertainty. Finally, we present step-by-step the new methodology and give a three-dimensional visualisation of the confidence intervals. We tackle the challenges of estimating parameters when the initial number of susceptibles and infected is unknown.

Maximum Likelihood (ML) is one of the most versatile analytic procedures for fitting statistical models to data, dating back to early works of Fisher around 1925. Typically, it finds parameter estimates that maximise the likelihood of a given dataset. There are many advantages of using likelihood-based approaches. Firstly, they are flexible, being applicable to a wide range of statistical models and various type of data sets (i.e. discrete, continuous, truncated, categorical, etc). Secondly, not only can they estimate parameters values, but also provide confidence intervals to characterise the uncertainty inherent in these estimates, due to their asymptotic normality propriety. Finally, they can be regarded as a unifying framework, as many common statistical approaches represent special cases of them. For example, Least Squares fitting is equivalent to Maximum Likelihood when the errors are normally distributed. To summarise, Maximum Likelihood based approaches are considered to be more robust, have better sufficiency and smaller errors than other methods.

4.1 Objective Function

Similar to Least Squares, the first step is defining an objective function. Maximum Likelihood finds a solution by maximising a likelihood function, defined as the probability of a given dataset having occurred, given a particular hypothesis. This is algebraically equivalent to Equation 4.1:

$$\mathcal{L}(D|H) = \mathcal{P}(D|H) \tag{4.1}$$

where D represents the observed data set and H is the hypothesis to be tested.

More precisely, the likelihood function is characterised by Equation 4.2.

$$\mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{4.2}$$

where x_i are the observed data points, θ is the vector of unknown parameters and $f(x_i, \theta)$ is the associated probability density function.

However, it is usually computationally more convenient to make use of the natural logarithm of the likelihood function, referred to as the *log likelihood*. Mathematically, this is defined in Equation 4.3.

$$\log \mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{4.3}$$

where x_i are the observed data points, θ is the vector of unknown parameters and $f(x_i|\theta)$ is the associated probability density function, as before.

This substitution is possible due to the increasing monotonicity of the logarithm function. This property makes both the logarithm function and the function itself achieve the maximum value at the same points. There are two main computational advantages of using the logarithm of the function. Firstly, the natural logarithm reduces the potential for underflow that may be caused by very small likelihoods. The second advantage arises when computing the derivative of the function, which is required to find its maximum. The likelihood function factorises into a product of functions, as shown in Equation 4.2, because the observed data points are assumed to be independent of each other. However, the logarithm of this product becomes a sum of individual functions in Equation 4.3, which is considerably easier to differentiate than a product.

In our implementation, we minimise the *negative log likelihood* function instead, as defined by Equation 4.4, which is just an equivalent characterisation.

$$\text{neg}\log \mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = -\sum_{i=1}^{n} \log f(x_i|\theta) \tag{4.4}$$

Based on the observation in Equation 4.5,

$$\underset{x}{\text{argmax}}(x) = \underset{x}{\text{argmin}}(-x) \tag{4.5}$$

the equivalence in Equation 4.6 holds.

$$\underset{x}{\text{argmax}} \sum_{i=1}^{n} \log f(x_i|\theta) = \underset{x}{\text{argmin}}(-\sum_{i=1}^{n} \log f(x_i|\theta)) \tag{4.6}$$

Figure 4.1 illustrates the R implementation for our objective function, that minimises the negative log-likelihood of the observations given any parameterisation of the model. Similar to the Least Squares approach, we prevent our optimisation procedure from searching over the entire range between $-\infty$ to $+\infty$ by using a *log transformation* on the parameters. Consequently, we constrain the search space to 0 to $+\infty$, which is the meaningful one in the context of our model.

```
sir.nll <- function(params, data){
  logBeta  <- params[1]
  logGamma <- params[2]

  pred <- as.data.frame(ode(y=c(S=500,I=10,R=0),
                            times=data$Day,
                            sir.model,
                            parms=exp(params)))
  logDensities <- dpois(x=data$I, lambda=pred$I, log=TRUE)
  nll <- -sum(logDensities)
}
```

FIGURE 4.1: R implementation of the negative log-likelihood function.

Note that we assume the observations to be *Poisson distributed*, making use of the function *dpois* in the R package *stats* that returns the log density. According to standard texts, epidemiologists model variability in disease occurrence using either the binomial, the Poisson or the exponential distribution. The authors of [25] argue that the three distributions have common attributes that lead to similar results for modelling variance in disease occurrence. They also state that the Poisson distribution is widely used by epidemiologists when the data involves counts of cases. Moreover, since we deal with discrete observations, the variance is expected to scale with the number of infected individuals [26] [27].

4.2 Optimisation Technique

The second step computes parameter estimates that minimise the negative log likelihood function by taking its derivative. The idea of finding the maximum or the minimum of

a function by taking its derivative is based on the extreme value theorem. This states that if a function $f(x)$ is continuous on a closed interval $[a, b]$, then $f(x)$ has a maximum and minimum value on the interval $[a, b]$. Algebraically, there exist x_{min} and x_{max} such that the formula in Equation 4.7 holds.

$$\mathrm{f}(x_{min}) \leq \mathrm{f}(x) \leq \mathrm{f}(x_{max}), \forall x \in [a, b] \tag{4.7}$$

Besides this theorem, there are two additional observations to be made. Firstly, the slope of the tangent line of the maximum and minimum is 0. Secondly, after the maximum the function decreases, and after the minimum it increases in value. Hence, the following two conditions must be met by x_{max} in order for it to be a maximum of a function f: $\mathrm{f}'(x_{max}) = 0$ and $\mathrm{f}''(x_{max}) < 0$. Similarly, x_{min} must meet the following two conditions in order to be a minimum of a function f: $\mathrm{f}'(x_{min}) = 0$ and $\mathrm{f}''(x_{min}) > 0$.

For multiple unknown parameters θ_i, finding Maximum Likelihood based estimates becomes more challenging. The estimation requires determining the simultaneous solution set for k equations, where k in the number of unknowns. Particularly, for the negative log likelihood function $\mathrm{neg}\log\mathcal{L}$ and $k = 2$, the system is shown in Equation 4.8.

$$\begin{array}{rcl} \frac{\partial \mathrm{neg}\log\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_1} & = & 0 \\ \frac{\partial \mathrm{neg}\log\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_2} & = & 0 \end{array} \tag{4.8}$$

In our implementation, we achieve this through the *mle2* function in the *bbmle* R package, which provides tools for general maximum likelihood estimation. This function uses the same optimiser that we used for Least Squares, *optim* from the *stats* package, which is based on the *Nelder-Mead algorithm*. It also computes an approximate covariance matrix for the parameters by inverting the Hessian matrix at the optimum, which can be later used to derive confidence intervals.

Figure 4.2 illustrates an example of curve fitting using the Maximum Likelihood based approach described. The observed data was generated synthetically with parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S0 = 500$, $I0 = 10$, $R0 = 0$ over a period of 100 days, within a closed population.

4.3 Confidence Intervals

A confidence interval statistically measures the reliability of an estimate. It aims to answer questions related to how to deal with uncertainty surrounding estimates, especially if they are derived from data that only represent a subset of the total population.

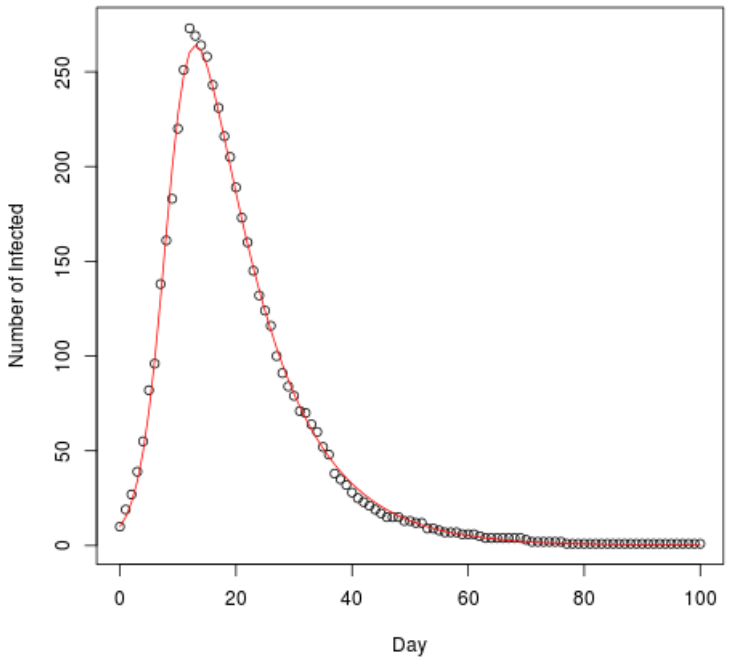


FIGURE 4.2: Curve fitting using Maximum Likelihood.
Original parameter values: $\beta = 0.001$, $\gamma = 0.1$.
Estimated parameter values: $\beta = 0.001034596$, $\gamma = 0.092604365$.

The interpretation of a confidence intervals is not strictly a mathematical issue, but also a philosophical matter [28]. Mathematics has only a limited role in deciding why an approach is preferred to another. Generally, there are multiple interpretations that can be given to a confidence interval. For the purpose of our work, we will consider the case expressed in terms of repeated samples. The 95% confidence interval will ideally contain the true value of the parameter 95% of the time, given repeated fittings of the model. It is only by chance that the true value of the parameter lies outside the confidence interval with probability 5%.

Traditionally, the Wald-type confidence intervals are widely used as an approximation to profile intervals. The standard procedure for computing such a confidence interval is by applying Equation 4.9.

$$estimate \pm (percentile \times SE(estimate)) \tag{4.9}$$

where SE is the standard error and the percentile is selected according to the desired confidence level and a reference distribution, i.e. a t-distribution for regression coefficients in a linear model, otherwise a standard normal distribution.

They are easier to compute for complex models, but perform poorly when the likelihood

surface is not quadratic. Additionally, a markedly skewed distribution of the parameter estimator or a standard error that poorly approximates the standard deviation of the estimator may affect their performance. Moreover, for generalised linear models, the standard errors are based on asymptotic variance derived from the covariance matrix. For small to medium sample sizes, this scenario may also cause poor performance.

Profile likelihood confidence intervals represent a more robust approach [29]. They do not assume normality of the estimator and appear to perform better for small samples sizes than Wald-type confidence intervals. These confidence intervals are based on an asymptotic approximation of the χ^2 distribution of the log-likelihood ratio statistic. To define them, we consider a model with θ the parameter of interest and δ a vector of the other parameters, and a likelihood function $\mathcal{L}(\theta, \delta)$. Then, the profile likelihood function for θ, $\mathcal{L}_1$ is given by Equation 4.10. By definition, the profile likelihood equals the maximum value of the likelihood function for every point.

$$\mathcal{L}_1(\theta) = \max_{\delta} \mathcal{L}(\theta, \delta) \tag{4.10}$$

The $100(1-\alpha)$ profile confidence interval is computed by inverting the likelihood ratio test. Mathematically speaking, given a parameter θ, it contains the set of all values θ_0 that do not reject the two-sided test of the null hypothesis $H_0 : \theta = \theta_0$ at a level of significance α. The likelihood ratio test statistic of the hypothesis is defined by:

$$D = 2\ [log\mathcal{L}(\hat{\theta}, \hat{\delta}) - log\mathcal{L}_1(\theta_0)] \tag{4.11}$$

Based on the asymptotic χ^2 distribution of the likelihood ratio test statistic, if the null hypothesis is true, the test of $H_0 : \theta = \theta_0$ will not be rejected at the level of significance α if and only if the relation in Equation 4.12 holds.

$$D \leq \chi^2_{1-\alpha} \tag{4.12}$$

where $\chi^2_{1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2 distribution on 1 degree of freedom.

We compute two sided confidence intervals using the *confint* function in the *bbmle* R package, as illustrated in Figure 4.3. Figure 4.4 is its corresponding two-dimensional contour plot, generated using function *curve3d* in the R package *emdbook*, and *contour*, *points* in the package *graphics*.

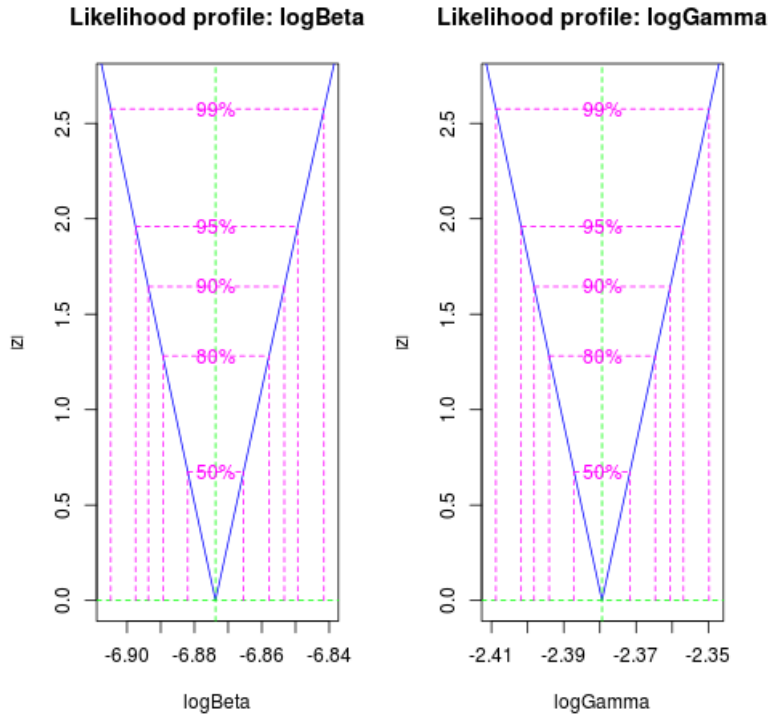


FIGURE 4.3: Profile confidence intervals for parameters $\log(\beta)$ and $\log(\gamma)$ estimated by Maximum Likelihood for the SIR curve fitted in Figure 4.2.

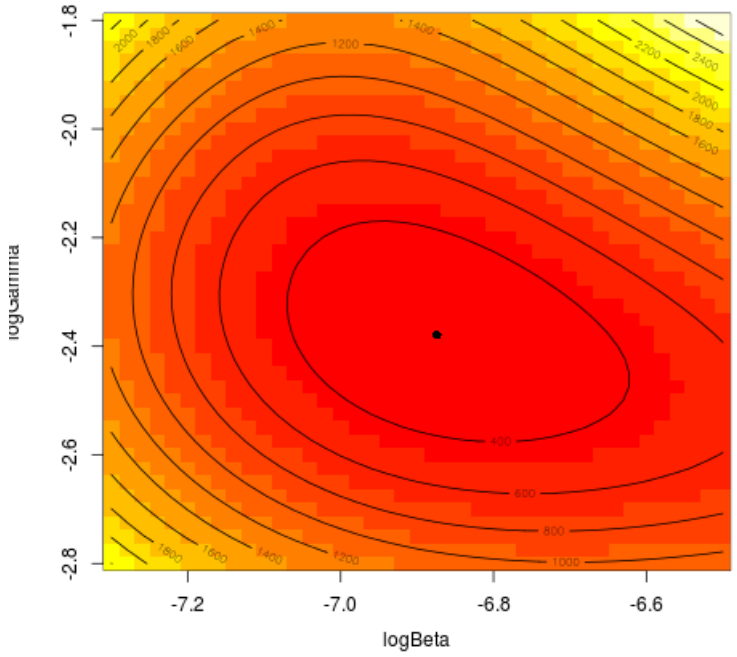


FIGURE 4.4: Contour plot for parameters $\log(\beta)$ and $\log(\gamma)$ estimated by Maximum Likelihood for the SIR curve fitted in Figure 4.2.

4.4 New Approach for Uncertainty Characterisation

We implemented a generic Maximum Likelihood based methodology for on-the-fly epidemic fitting of SIR model from a single trace, which yields confidence intervals on parameter values. The method is fully automated and avoids the laborious manual efforts traditionally deployed in the modelling of biological epidemics.

Mathematical modelling of infectious disease dynamics relies on a series of assumptions regarding key parameters that cannot be measured directly. We consider the challenges of estimating the initial number of susceptible and infected individuals in the target population, when these values are unknown. Currently, there is no principled way of doing this, as traditionally they are either known or can be estimated from the context. However, in an era of social and technological epidemics, we argue that time and speed of movement make it infeasible to provide accurate estimates within a reasonable time frame that allows quick action to be taken. Based on the SIR model, we developed two new methodologies, one for estimating the vector of parameters β, γ, S_0, and one for β, γ, S_0 *and* I_0.

4.4.1 Data Transformation

By definition, all SIR model parameters we are interested in estimating represent positive quantities. This observation allows us to apply data transformation techniques in order to prevent the optimisation from exploring infeasible values. The obvious choice is to use a log transformation, as before. This prevents the optimisation procedure from searching over the entire range from $-\infty$ to $+\infty$, being instead constraint between 0 and $+\infty$.

Another key observation is that the initial number of infected individuals is always smaller than the initial number of susceptibles. This allows us to apply a logistic-based transformation, reducing the optimisation search space of I_0 between 0 and S_0. A similar transformation can be applied to the initial number of susceptibles S_0 when the target population is bounded by a known value.

A *logistic* function is a special case of a sigmoid function, often used to model population growth. It is defined for real values of x from $-\infty$ to $+\infty$, taking values within the range (0, 1), as shown by Equation 4.13.

$$\text{logistic}(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{4.13}$$

The inverse of the logistic function is the natural *logit* function, being often used in statistics for parameter representing probabilities. It is defined for parameters p between

0 and 1, and takes values from $-\infty$ to $+\infty$, as shown in Equation 4.14.

$$\text{logit}(p) = \log(\frac{p}{1-p}) \tag{4.14}$$

We modify the logistic function and its inverse in order to map the searching space ranging from $-\infty$ to $+\infty$ to one between 0 and $maxValue$, where $maxValue$ represents either the number of initial susceptibles, or the population bound, as discussed above. Hence, the transformation function is given by Equation 4.15

$$\text{trans}(x) = \frac{maxVal}{1+e^{-x}} \tag{4.15}$$

and its corresponding inverse is defined in Equation 4.16.

$$\text{untrans}(y) = \log(\frac{y}{maxVal-y}) \tag{4.16}$$

4.4.2 Parameter Space Searching

To account for uncertainty as each outbreak unfolds over time, we apply our fitting methodology on *truncated data sets*. For each dataset, we initially consider the first 3 observations. Then, we add one observation at a time, until we reach the end of the outbreak, creating new truncated datasets each time.

We compute parameters estimations for each truncated dataset, considering in turn each of the following vectors of unknown parameters: β, γ, S_0 and β, γ, S_0, I_0. The process of searching the parameter space for each set of parameters takes place in two stages. First, we find the parameters estimates that give the best fit to the data based on a Least Squares objective function. In order to avoid the optimisation procedure being trapped in a *local minimum*, we restart for 20 different randomly chosen initial values of the parameter vectors. Sensible restrictions are imposed, such as $0 < \beta < \gamma$ and $0 < I_0 < S_0$. The final candidate is selected to be the vector that yields the lowest value for the initial number of susceptibles S_0 across all runs. Then, we use the parameter estimates obtained in the first stage as initial values for a Maximum Likelihood fitting procedure. The reason behind this approach is to overcome *computational challenges* that arises through the estimation and confidence interval calculation within the *mle2* function in the R package *bbmle*. Maximum Likelihood based approaches are considerably more sensitive to the initial guesses of the parameters to be estimated than the Least Square ones. If these guesses are not within a sensible range, being too far off the true values, then the output becomes unreliable. Moreover, computing the confidence intervals using *mle2* involves

calculating the covariance matrix for the parameters, which is done by inversion of the Hessian matrix at the optimum. This procedure can also be unsuccessful depending on the initial parameters.

4.4.3 Visualisation

Once the fitting procedure completes successfully, we provide a three-dimensional visualisation of the confidence intervals when the vector of unknown estimates is β, γ, S_0. It expands on the idea of a two-dimensional contour plot previously explored in the literature, where each contour line connects values that lie within the same confidence interval. We take this approach one step further and add more dimensions. The resulting representation is therefore based on isosurfaces, the three-dimensional analog of isolines. Each of the surfaces represent parameters values that lie in the same confidence interval within a volume of space. Their shape is a variation of an ellipsoid. The three unequal axes represent the size of the confidence intervals corresponding to each of the three parameters. These intervals may be asymmetrical, hence the surfaces are described by a modified version of the ellipsoid's in Equation 4.17. Higher dimensional analog representation require more complicated equations and are difficult to visualise, hence are not discussed further.

$$\frac{(x-x_0)^2}{a_1 a_2} + \frac{(y-y_0)^2}{b_1 b_2} + \frac{(z-z_0)^2}{c_1 c_2} = 1 \tag{4.17}$$

where x_0, y_0, z_0 are the coordinates of the origin - the true values in our case, and a_i, b_i, c_i are the corresponding confidence intervals sizes.

The isosurface plots were generated in Matlab. A sample example is illustrated in Figure 4.5. It clearly indicates that the estimated range of possible values is wider as the confidence level increase.

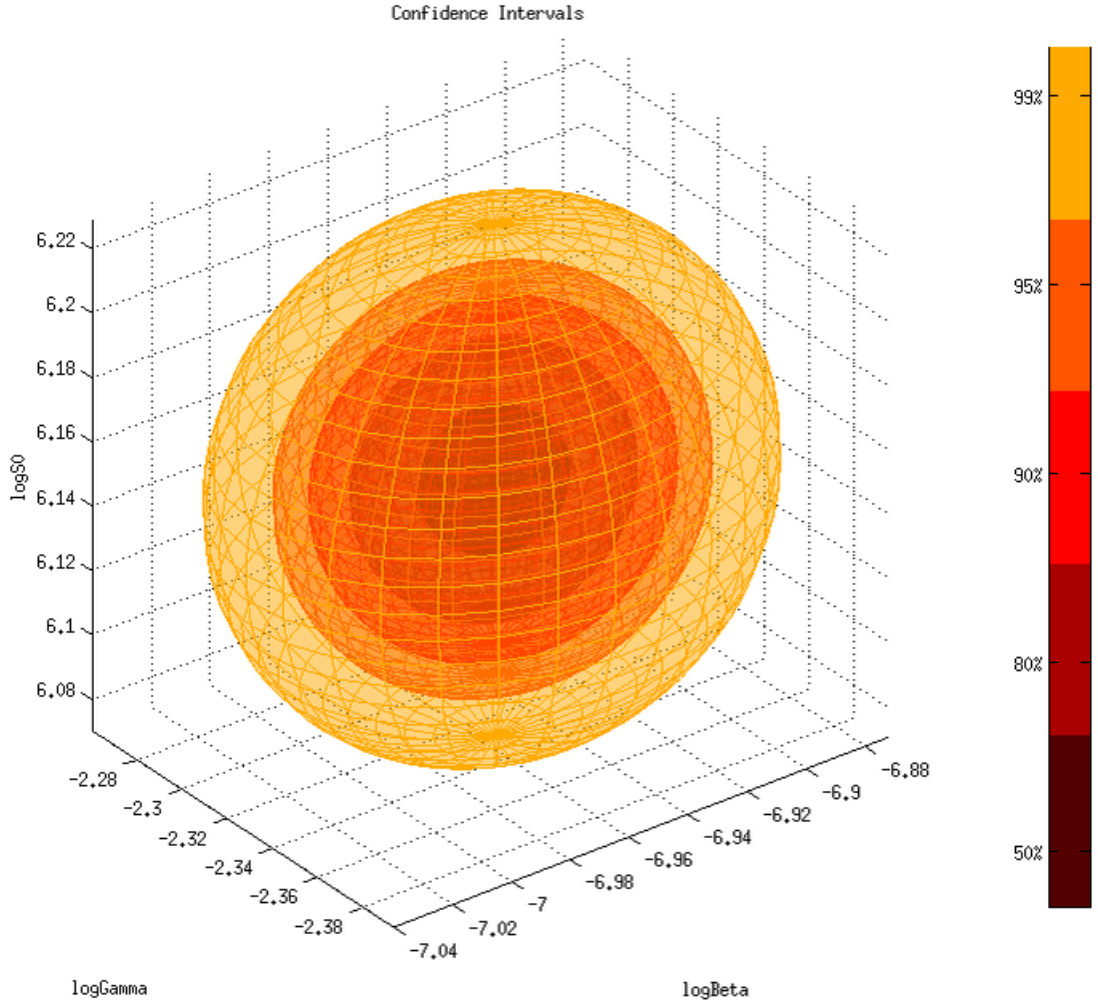


FIGURE 4.5: Isosurface representing the profile confidence intervals of parameters $\log(\beta)$, $\log(\gamma)$ and $\log(S_0)$ for data fitting on SIR model

Chapter 5

Evaluation

This section presents key results in validating our on-the-fly epidemic fitting methodologies on both synthetic and real datasets. For the Least Squares based approach, we use the coefficient of determination as a metric in assessing the efficacy of the method. For the Maximum Likelihood based approach, we analyse the confidence intervals of the estimates and measure the true value recoverability rate.

Because we fit our datasets as the epidemic unfolds over time, our methodologies are applied on truncated datasets. For each dataset, we initially consider the first 3 observations. Then, we add one observation at a time, until the end of the outbreak, creating new truncated datasets each time. We estimate parameters for each truncated dataset.

5.1 Synthetic Data

The synthetic datasets were generated based on Gillespie's Stochastic Simulation Algorithm, using the *ssa* function in the *GillespieSSA* R package. The use of synthetic datasets allows us to evaluate the ability of our methodology to recover model parameters from a single trace, as the ground truth is known.

One very important aspect we had in mind when generating the synthetic data was to avoid the so called "inversion crime" phenomenon. This expression refers to the act of using the same model to both generate and invert synthetic data. The stochastic nature of our data generation process avoids such issues.

We ran our experiments on 1000 different synthetic datasets. The one used for illustrative purposes in the rest of this section was generated by simulating a SIR epidemic with known parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S_0 = 500$, $I_0 = 10$, $R_0 = 0$, in a closed population, over a period of 100 days. It is depicted in Figure 5.1.

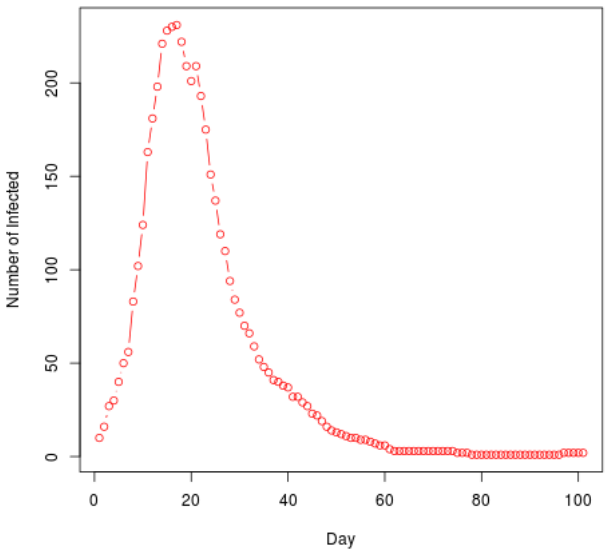


FIGURE 5.1: Synthetic dataset for a SIR epidemic with parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S_0 = 500$, $I_0 = 10$, $R_0 = 0$

5.1.1 Fitting Using Least Squares

In this section we discuss the results of applying our Least Squares based fitting procedure of truncated synthetic datasets on SIR model, as the epidemic unfolds over time. We consider two methodologies, one to estimate the vector of unknown parameters β, γ, illustrated in Figure 5.2, and one to estimate β, γ, S_0, as shown in Figure 5.3.

From a very early stage, and taking into account only the first 10 observations, our model predicts with surprising precision the peak of the epidemic in both cases. As time progresses, the fits become more and more stable and closer to the original epidemic curve. We observe that the estimated parameters for the best fit curve are very close to their true values, and the predicted curves fit well the data points. The coefficient of determination R^2 gets closer to 1 as time progresses, indicating that new observations improve the fit, as expected.

When comparing the two procedures, we notice that adding more parameters to the set of unknowns improves the overall quality of the fits. When we assume the initial number of susceptibles S_0 unknown, and take only the first 10 observations, the predicted curve does not fit the data points as well as when fixing S_0, but it predicts with higher accuracy the peak of the epidemic. In addition, the prediction with S_0 unknown produce higher values for the coefficient of determination R^2 once the model becomes stable - 0.9956 as compared to 0.9935. This proves the importance of accounting for parameter uncertainty in model-based analysis.

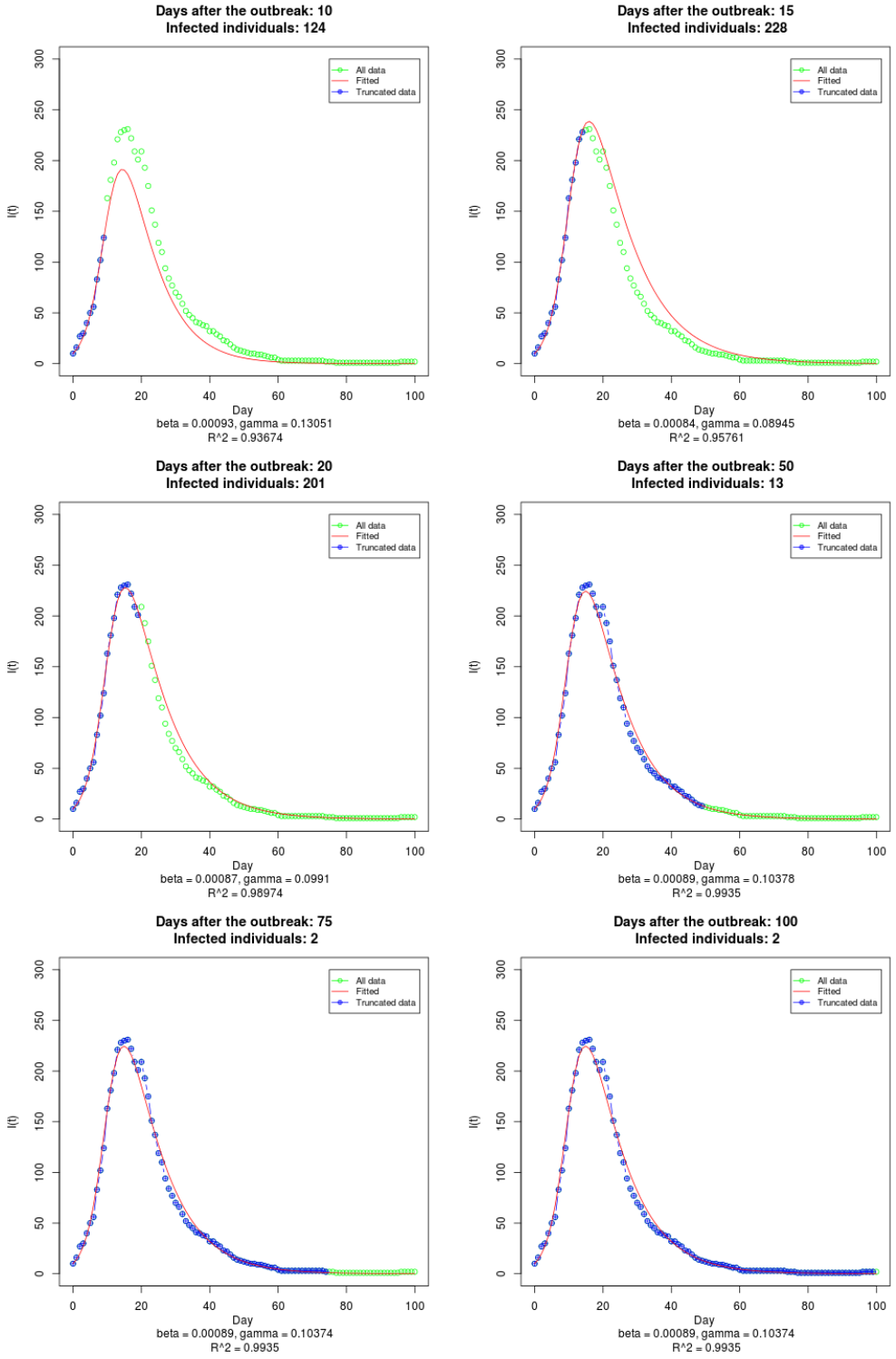


FIGURE 5.2: Over-time Least Squares fitting of synthetic data on SIR model with unknown parameters β and γ.

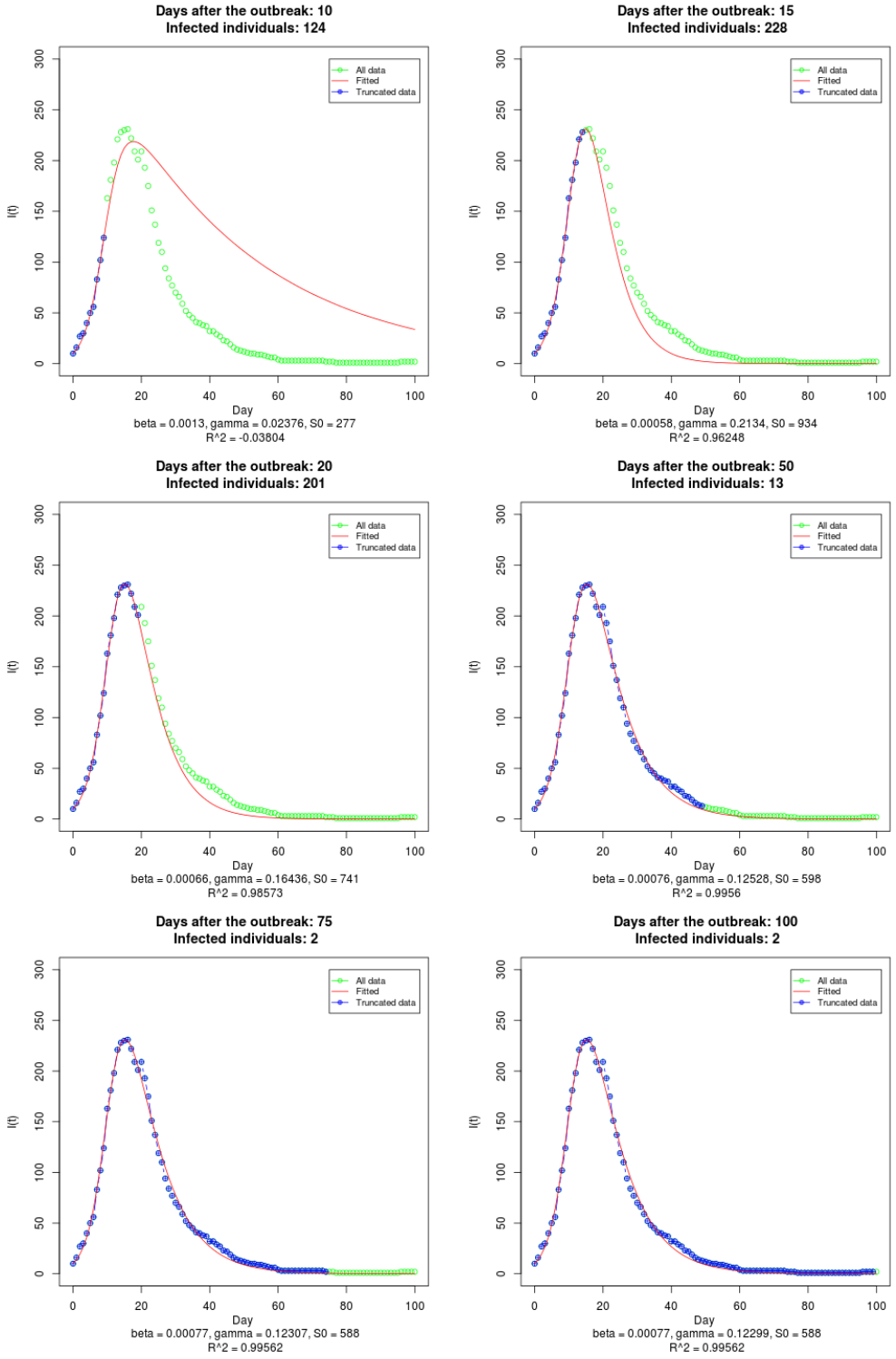


FIGURE 5.3: Over-time Least Squares fitting of synthetic data on SIR model with unknown parameters β, γ and S_0.

5.1.2 Fitting Using Maximum Likelihood

In this section we discuss the results of applying our Maximum Likelihood based fitting procedure of truncated synthetic datasets on SIR model, as the epidemic unfolds over time. We consider three methodologies, one to estimate the vector of unknown parameters β, γ, illustrated in Figure 5.4, one to estimate β, γ, S_0, shown in Figure 5.7 and one to estimate β, γ, S_0, I_0, depicted in Figure 5.10. We also discuss below the uncertainty characterisation for each of these methodologies.

From a very early stage, and taking into account only the first 15 observations, our model predicts with surprising precision the peak of the epidemic in all three cases. As time progresses, the fits become more and more stable and closer to the original epidemic curve. We observe that the estimated parameters for the best fit curve are very close to their true values, and the predicted curves fit well the data points. The coefficient of determination R^2 gets closer to 1 as time progresses, indicating how new observations improve the fit.

When considering the vector of unknown parameters β, γ, we observe that the model prediction do not vary very much when new observations are added. For example, at day 15 $\beta = 0.00084$, $\gamma = 0.09064$, and at day 100 $\beta = 0.00089$, $\gamma = 0.10237$. However, adding S_0 as an unknown makes β vary between 0.0007 and 0.00095, and γ from 0.06095 to 0.015128. It is interesting to notice that in this case β, γ and S_0 do not progressively get closer to the true values as more observations are added, but rather arbitrarily increase and decrease in value, although the curve fittings improve over time. At day 100, their estimates are reasonably close to the true values. We suggest this is a direct consequence of allowing S_0 to vary, but keeping I_0 fixed. Our hypothesis is confirmed by the results obtained when adding I_0 as unknown. Accounting for the uncertainty inherent in I_0 leads to a progressive improvement in the quality of the predictions, with recovered parameters very close to the true values at the end of the epidemic: $\beta = 0.00099$, $\gamma = 0.09614$, $S_0 = 462$, $I_0 = 9$. Although the variance of I_0 is very small, as it is bound by the value of S_0, it does bring a significant improvement to our fitting procedure.

Figures 5.5, 5.8 and 5.11 illustrate the corresponding profile confidence intervals for fitting the synthetic dataset on SIR model with two, three, and four unknown parameters respectively.

We observed that the confidence intervals become narrower as more observations are added, indicating that the uncertainty in the parameters decreases. Table 5.1 shows some observations of lower and upper bounds on each parameter when the data is fitted over time.

Data%	β		γ		$S0$	
	Lower	Upper	Lower	Upper	Lower	Upper
25%	5.66e-04	8.47e-04	1.08e-01	1.93e-01	569	962
50%	7.17e-04	8.36e-04	1.17e-01	1.35e-01	590	692
75%	7.62e-04	8.68e-04	1.13e-01	1.26e-01	568	646
100%	8.39e-04	9.47e-04	1.03e-01	1.14e-01	519	582

TABLE 5.1: Confidence Intervals for over-time fitting of synthetic data on SIR model with unknwon parameters β, γ, S_0.

Figures 5.6 and 5.9 provide a graphical characterisation of the uncertainty when fitting the synthetic dataset on SIR model with two and three unknown parameters respectively. The first one is a two-dimensional contour plot; the contour lines connect values that lie within the same confidence interval. The second one is a three-dimensional isosurface plot; each surface embodies parameter values that lie in the same confidence interval within a volume of space. Both these diagrams highlight the fact that the estimated range of possible values is wider as the confidence level increase.

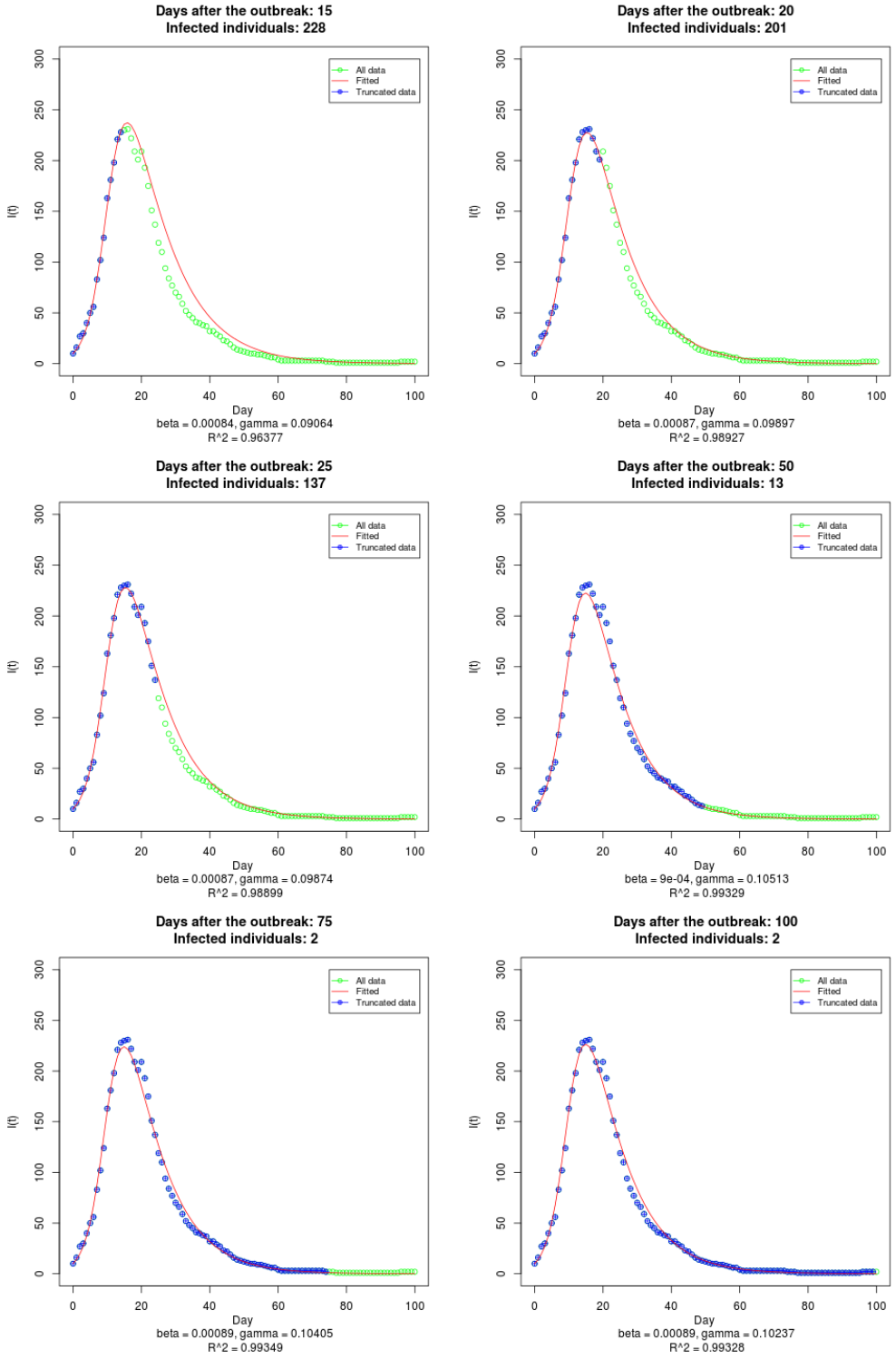


FIGURE 5.4: Over-time Maximum Likelihood fitting of synthetic data on SIR model with unknown parameters β and γ.

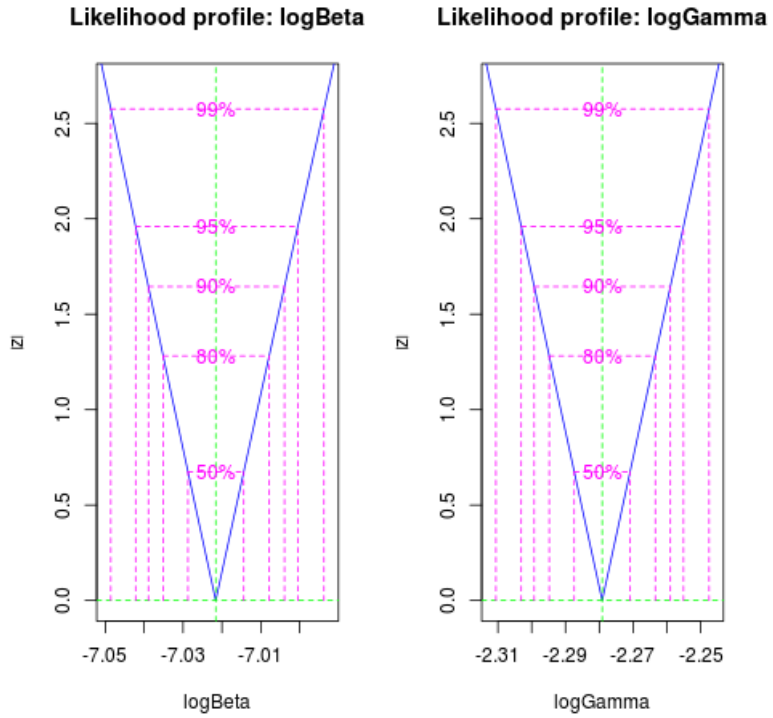


FIGURE 5.5: Likelihood profiles of parameters $\log(\beta)$ and $\log(\gamma)$ for Maximum Likelihood fitting of synthetic data on SIR model

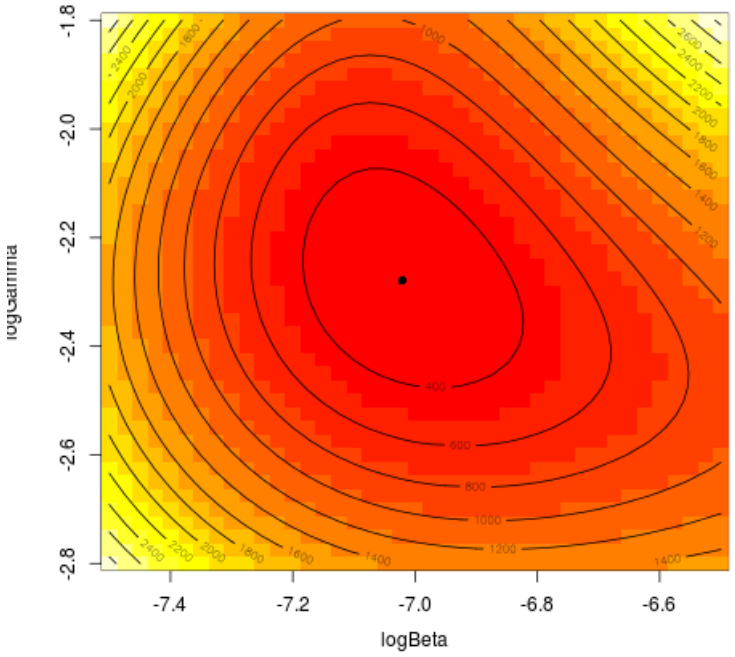


FIGURE 5.6: Contour plot of parameters $\log(\beta)$ and $\log(\gamma)$ for Maximum Likelihood fitting of synthetic data on SIR model

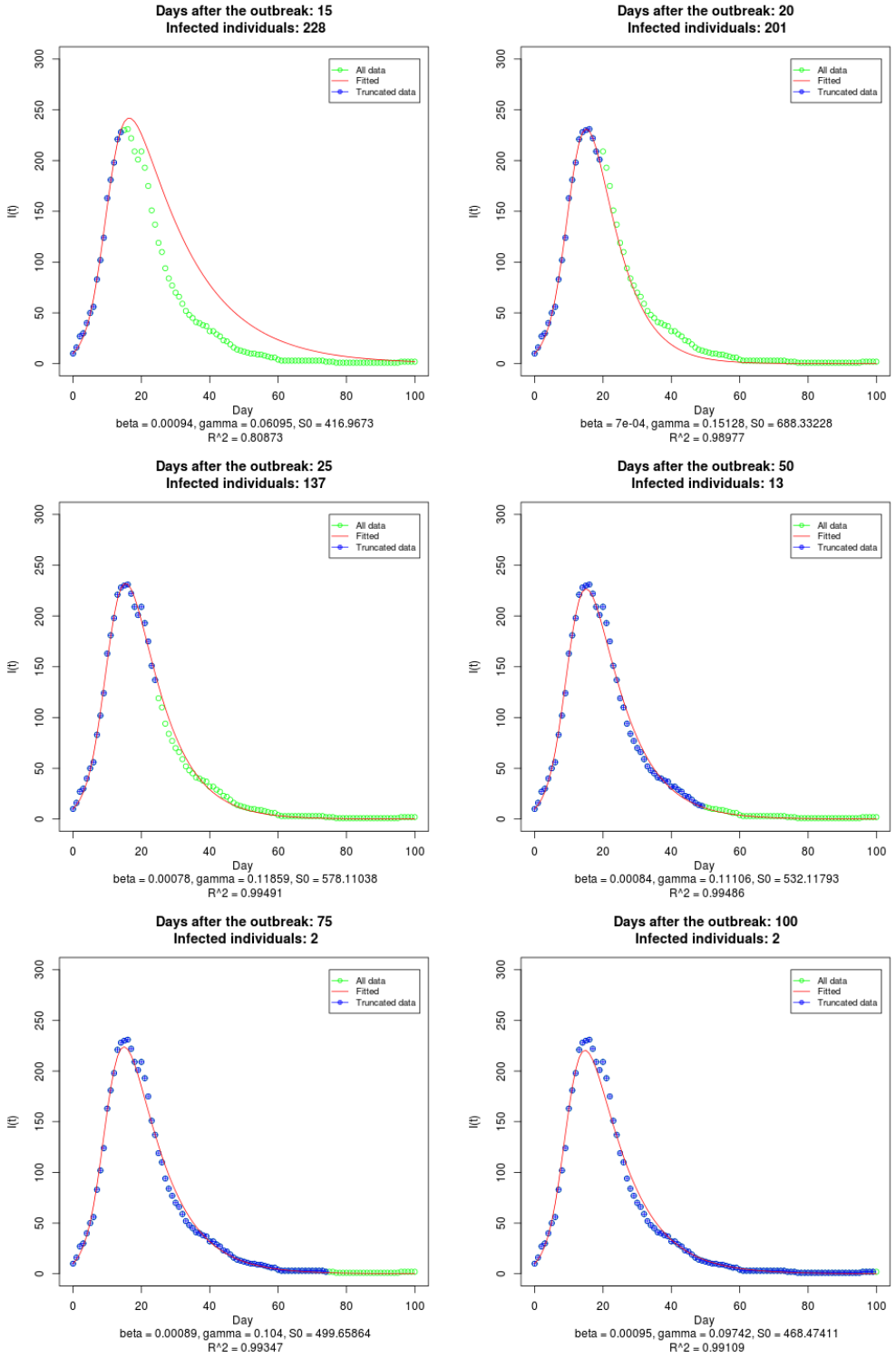


FIGURE 5.7: Over-time Maximum Likelihood fitting of synthetic data on SIR model with unknown parameters β, γ and S_0.

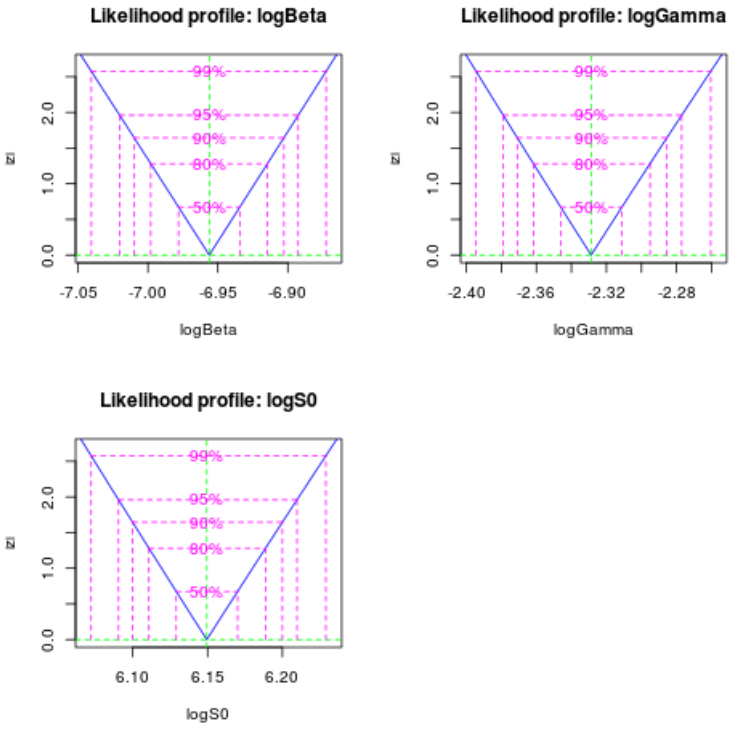


FIGURE 5.8: Likelihood profiles of parameters $\log(\beta)$, $\log(\gamma)$ and $\log(S_0)$ for Maximum Likelihood fitting of synthetic data on SIR model

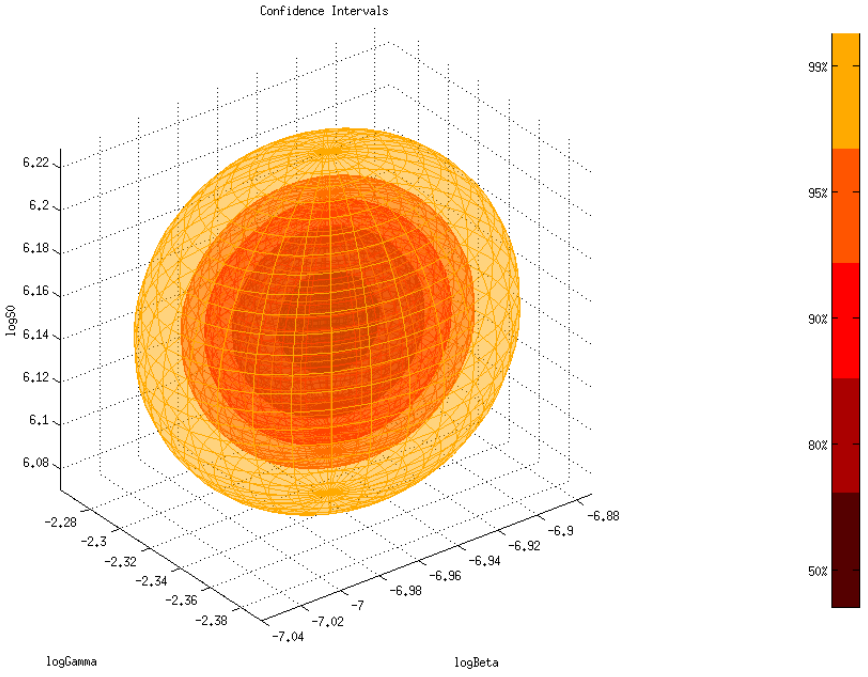


FIGURE 5.9: Isosurface plot of parameters $\log(\beta)$, $\log(\gamma)$ and $\log(S_0)$ for Maximum Likelihood fitting of synthetic data on SIR model

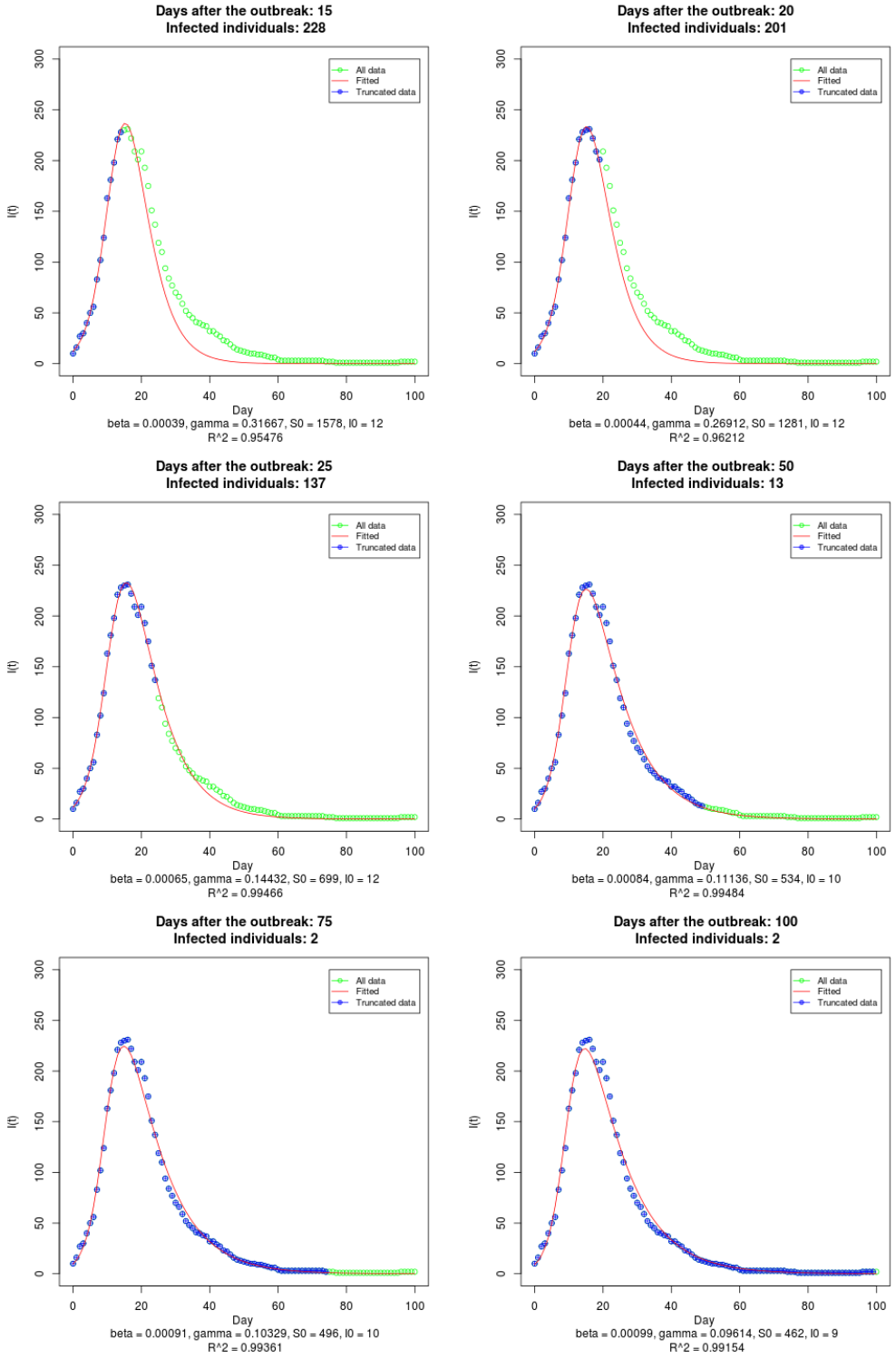


FIGURE 5.10: Over-time Maximum Likelihood fitting of synthetic data on SIR model with unknown parameters β, γ, S_0 and I_0.

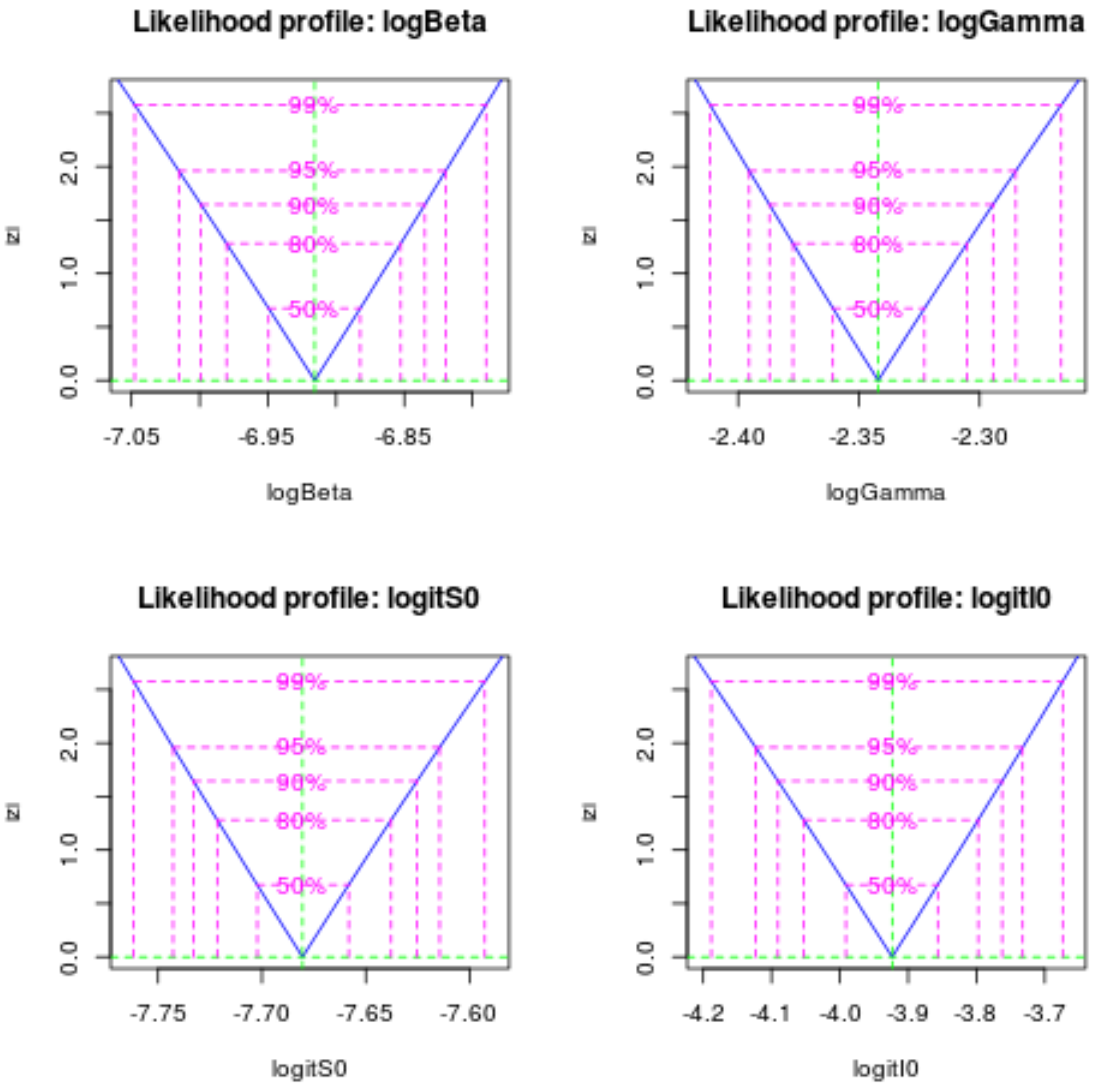


FIGURE 5.11: Likelihood profiles of parameters $\log(\beta)$, $\log(\gamma)$, $\log(S_0)$ and $\log(I_0)$ for Maximum Likelihood fitting of synthetic data on SIR model

True parameters recoverability rate

Another way of validating our methodology is by assessing the true value recoverability rate. We repeatedly run our fitting procedure over 1000 stochastic simulations of the SIR model, computing confidence interval for the estimated parameters each time. As vector of unknown parameters, we consider, in turn, β, γ, S_0 and β, γ, S_0, I_0.

One might hope that 95% of the times the true value will lie within the 95% confidence intervals. However, this is not the case, as illustrated in Tables 5.2 and 5.3. We notice that the recoverability rates for individual parameters vary between 26% and 48%, but

they become significantly lower when considering the recoverability rate for all parameters at once. This indicates that there are other sources of uncertainty that we have not considered in our model.

Another notable remark results from comparing the recoverability rate for the two vectors of unknown parameters: β, γ, S_0 and β, γ, S_0, I_0. When assuming I_0 fixed, the rate of recoverability for β is only 26.59%, but it almost doubles, increasing up to 41.99% when considering I_0 unknown. The rate for γ stays exactly the same, at 26.28%. For S_0 the rate increases slightly, from 31.82% to 34.44%. These results reveal that although the variance of I_0 is small, it has a high impact on the recoverability rate of true parameters and should therefore be considered for better predictions.

Parameter	Recoverability rate
β	26.59%
γ	26.28%
S_0	31.82%
β, γ, S_0	8.86%

TABLE 5.2: True value recoverability rate for unknown parameters β, γ and S_0 on synthetic data.

Parameter	Recoverability rate
β	41.99%
γ	26.28%
S_0	34.44%
I_0	48.04%
β, γ, S_0, I_0	9.46%

TABLE 5.3: True value recoverability rate for unknown parameters β, γ, S_0 and I_0 on synthetic data.

5.2 CDC Influenza Data

The real dataset represents positive Influenza cases summed over all subtypes of the flu virus, as reported to the Center of Disease Control and Prevention (CDC) during 2012 - 2013 Influenza season, starting in September 2012. This institution studies the impact of flu in the US. The data was obtained via the FluView Web Portal [1]. We chose Influenza because it is one of the most common infectious disease present in humans, with regular annual outbreaks. Hence, it allows us to evaluate the applicability of our methodology in real scenarios.

5.2.1 Fitting Using Least Squares

In this section we discuss the results of applying our Least Squares based fitting procedure of truncated real datasets on SIR model, as the epidemic unfolds over time. We consider two methodologies, one to estimate the vector of unknown parameters β, γ, illustrated in Figure 5.12, and one to estimate β, γ, S_0, as shown in Figure 5.13.

We manage to predict the peak in infectious individuals from only 9 observations to be around day 11 and of magnitude around 7000. In reality, it occurs to be only 1 day later, with approximately the same number of infected individuals. The accuracy of predicting from partial information on a single trace the time of the peak, the magnitude of the peak and the tail of the infection is remarkable. As time progresses, the fits become more and more stable and closer to the original epidemic curve. The coefficient of determination R^2 is very close to 1, stabilising it's value at 0.991 after

When comparing the two procedures, we notice that adding more parameters to the set of unknowns improves the overall quality of the fits. The prediction with S_0 unknown produces slightly higher values for the coefficient of determination R^2 once the model becomes stable - 0.99133 as compared to 0.99166. This confirms the results obtained on synthetic data, validating the applicability of our methodology to real world scenarios.

[1] http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

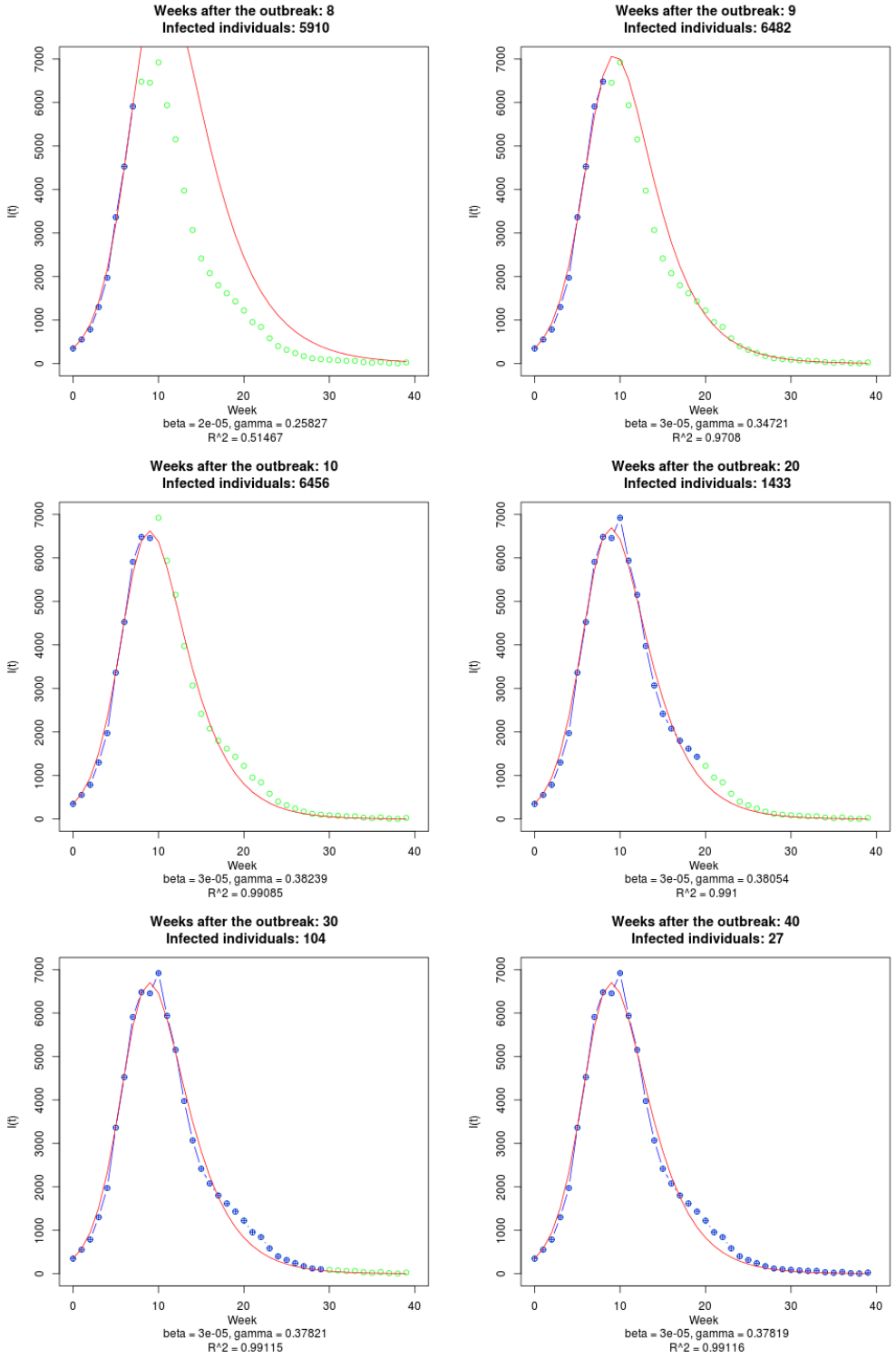


FIGURE 5.12: Least Squares fitting over-time of Influenza data on SIR model with β, γ unknown

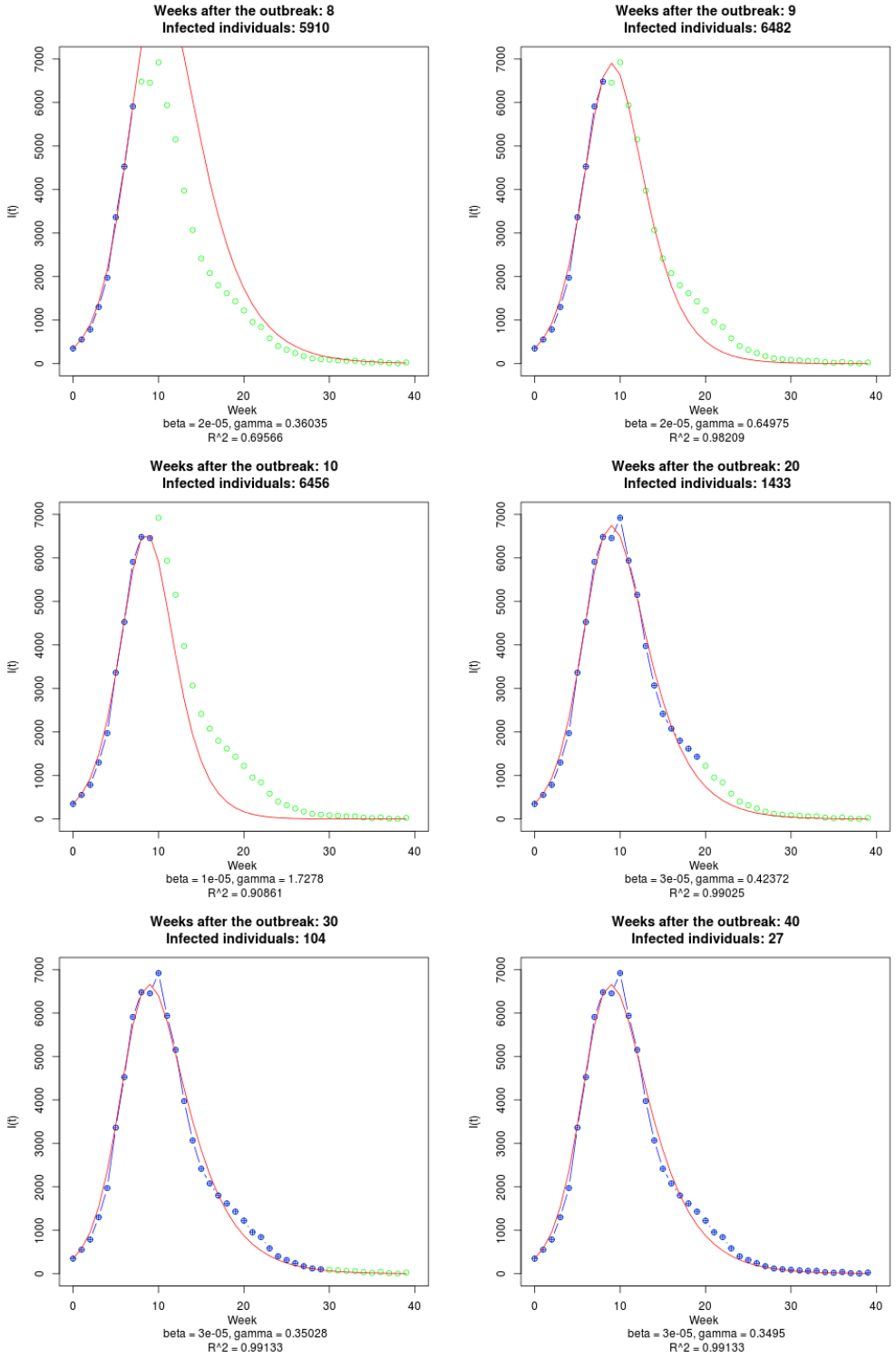


FIGURE 5.13: Least Squares fitting over-time of Influenza data on SIR model with β, γ, S_0 unknown

5.2.2 Fitting Using Maximum Likelihood

In this section we discuss the results of applying our Maximum Likelihood based fitting procedure of truncated real datasets on SIR model, as the epidemic unfolds over time. We consider two methodologies, one to estimate the vector of unknown parameters β, γ, S_0, shown in Figure 5.14, and one to estimate β, γ, S_0, I_0, illustrated in Figure 5.17. We also discuss below the uncertainty characterisation for each of these methodologies.

The experiments on real data confirmed the results obtained on synthetic datasets. The model proved highly accurate in predicting from partial information on a single trace. Considering only 20% of the observations, it accurately predicts the time of the peak, its magnitude, and the tail of the infection. As before, the confidence intervals for the parameter estimates become narrower as more observations are added, indicating that the uncertainty in the parameters decreases. Likelihood profiles, and corresponding iso-surface plot are illustrated in Figures 5.15, 5.16, 5.18. Table 5.4 shows some observations of lower and upper bounds on each parameter when the data is fitted over time.

TABLE 5.4: Confidence Intervals for over-time Influenza data (* - non convergence)

Data%	β		γ		$S0$	
	Lower	Upper	Lower	Upper	Lower	Upper
25%	*	*	*	*	*	*
50%	2.95e-05	3.22e-05	3.46e-01	3.81e-01	26769	30118
75%	3.50e-05	3.69e-05	2.90e-01	3.06e-01	22091	23515
100%	3.53e-05	3.70e-05	2.90e-01	3.03e-01	22031	23292

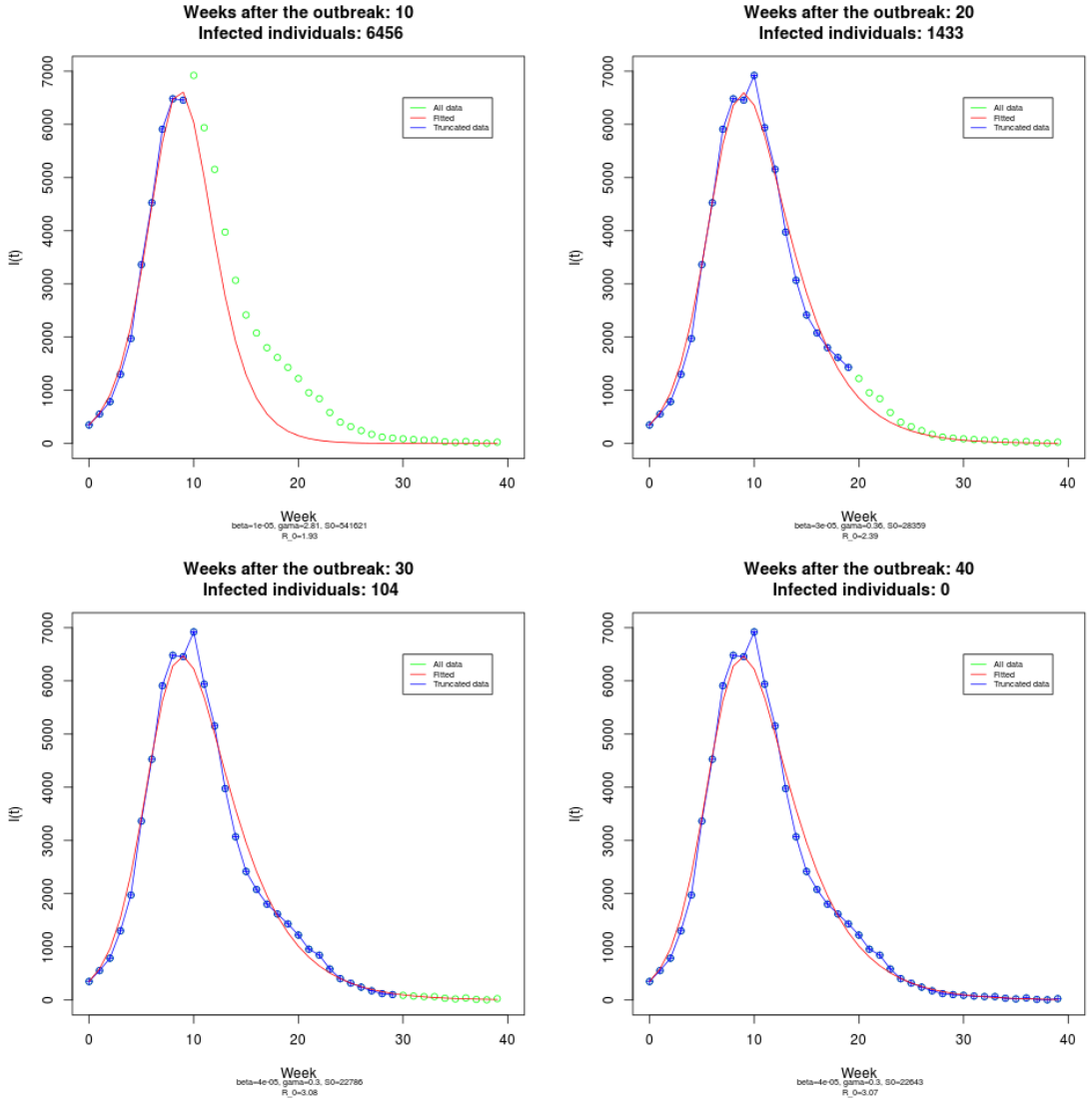


FIGURE 5.14: Over-time Maximum Likelihood fitting of Influenza data on SIR model with β, γ, S_0 unknown

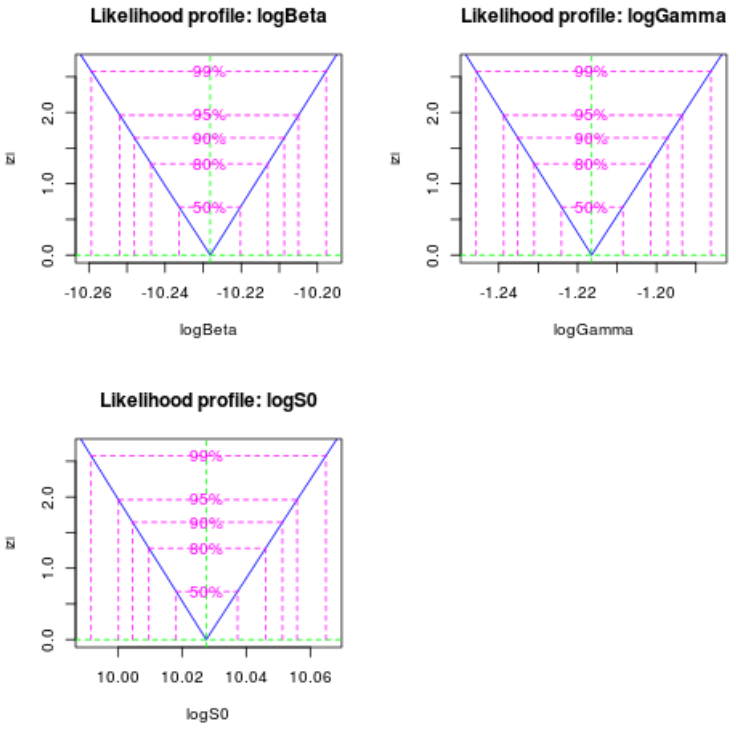


FIGURE 5.15: Likelihood Profile for Maximum Likelihood fitting of Influenza data on SIR model with β, γ, S_0 unknown.

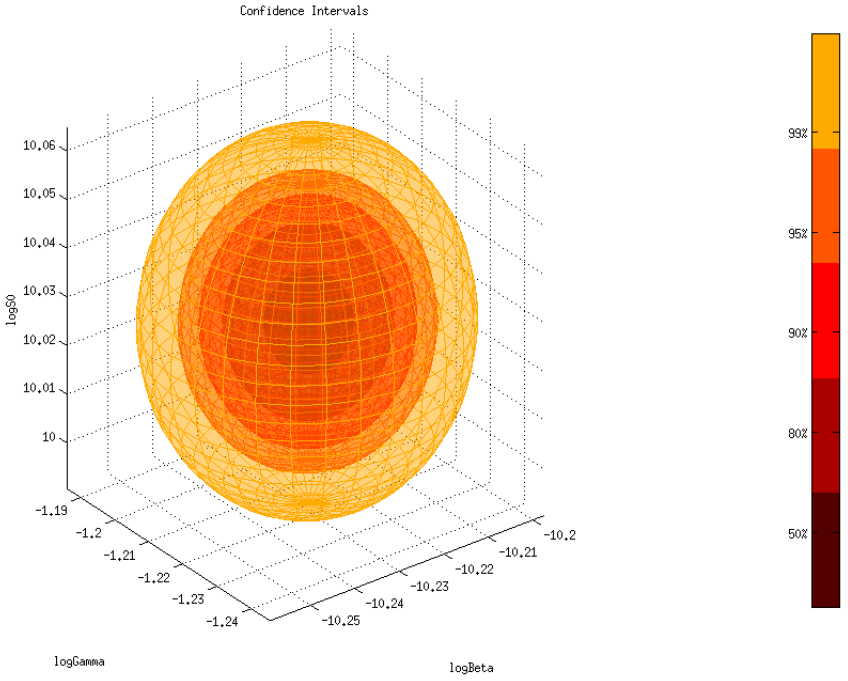


FIGURE 5.16: Isosurface plot of parameters $\log(\beta)$, $\log(\gamma)$ and $\log(S_0)$ for Maximum Likelihood fitting of Influenza data on SIR model

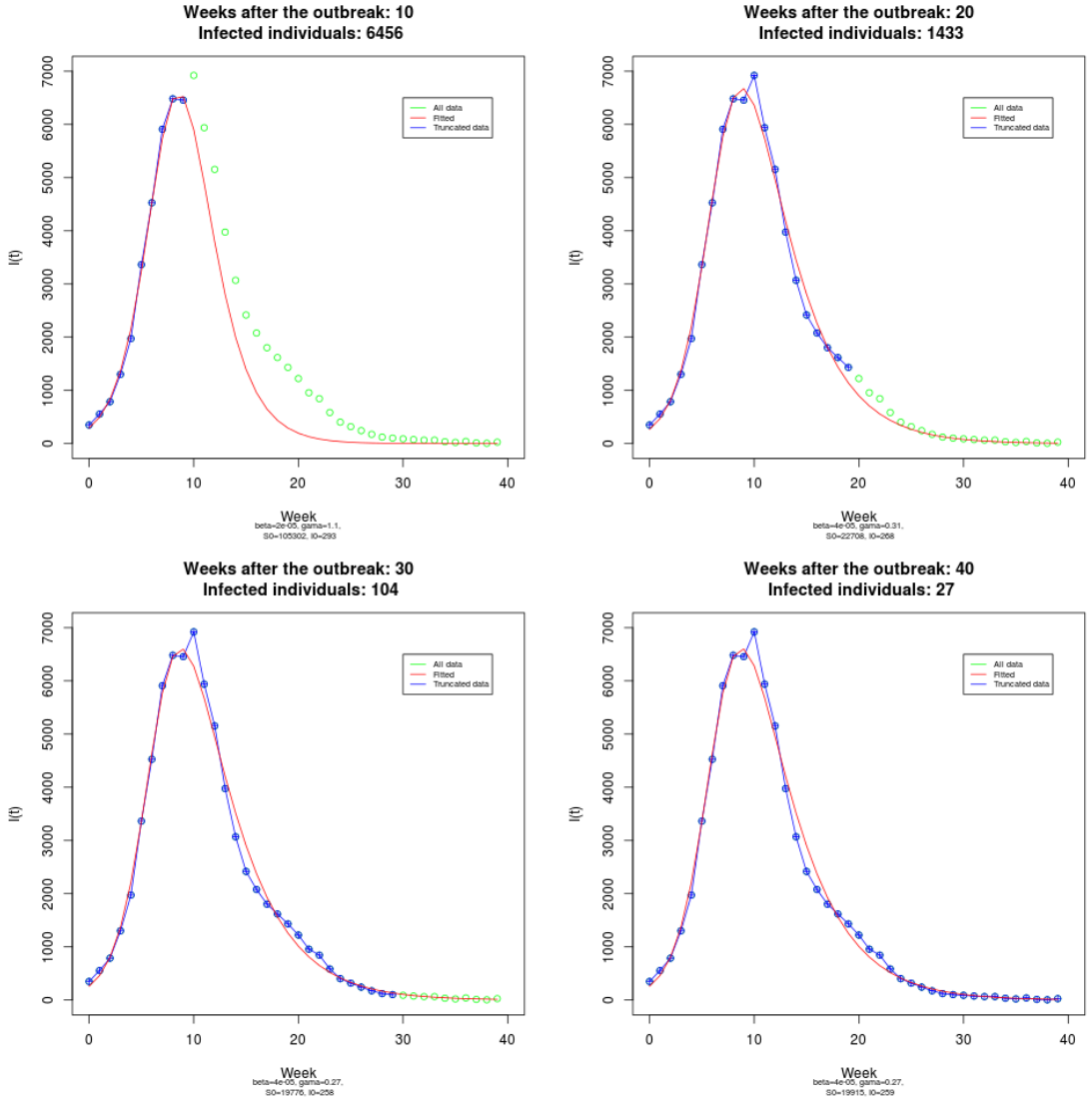


FIGURE 5.17: Over-time Maximum Likelihood fitting of Influenza data on SIR model with β, γ, S_0, I_0 unknown

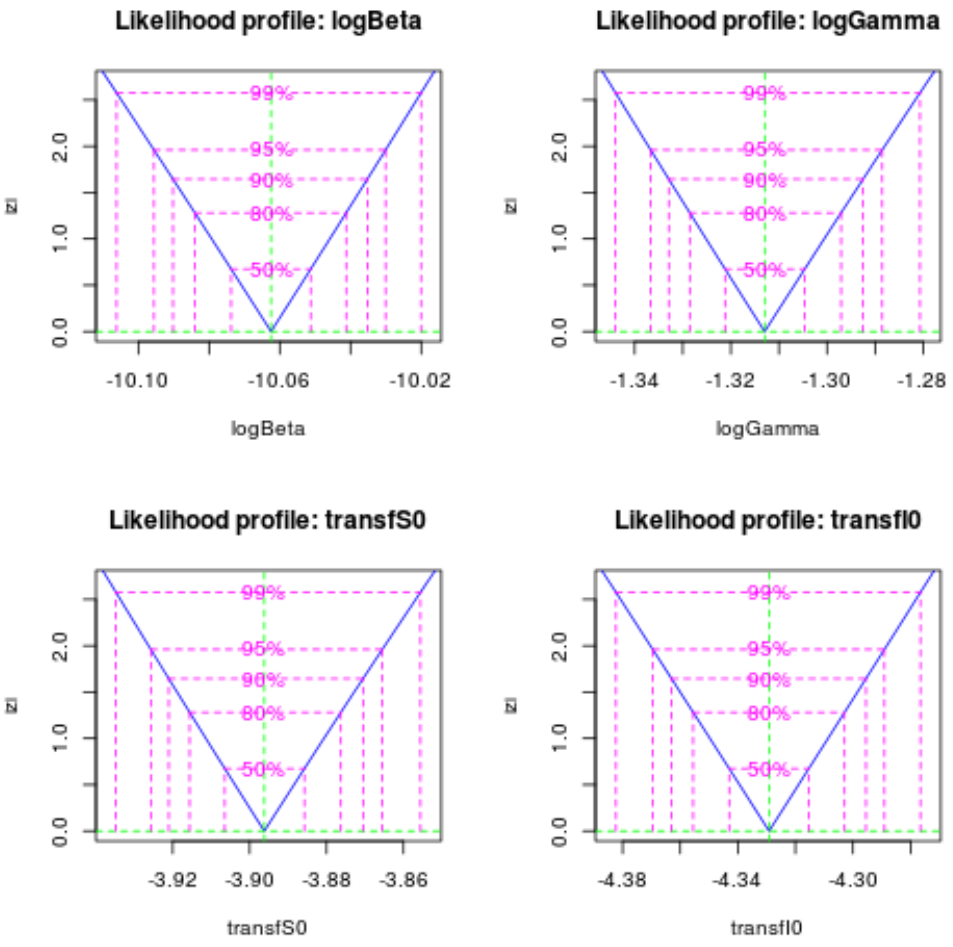


FIGURE 5.18: Likelihood Profile for Maximum Likelihood fitting of Influenza data on SIR model with β, γ, S_0, I_0 unknown

Chapter 6

Conclusion

6.1 Contributions

Researching applications of mathematical modelling techniques to epidemic phenomena proved to be a challenging, but rewarding task. While working on this project, we had the opportunity to deal with difficult aspects of statistical estimation of parameters and characterising their uncertainty. We tackled the challenge of estimating key parameters, such as the initial number of susceptible and infected individuals in the SIR model. Traditionally, they are assumed to be known or can be inferred from the context, but this approach is not feasible for modern outbreaks. Moreover, we had the pleasure to address unanswered questions of wide interest in areas such as contingency planning regarding the importance of uncertainty characterisation.

We initially implemented a Least Square based methodology for on-the-fly epidemic fitting on SIR models from a single trace. The method was validated using both synthetic and real data. From very early stages, our model predicted with surprising precision the peak of the epidemic. The estimated parameters for the best fit curve were very close to their true values, and the predicted curves fitted well the data points.

The main contribution of this project is a generic Maximum Likelihood based approach that characterises rigorously the uncertainty inherent in parameter estimates. It is addressed to on-the-fly epidemic fitting of SIR models from a single trace, and yields confidence intervals on parameter values. Opposite to traditional epidemiological modelling techniques, our approach is fully automated. We also provide estimates for key parameters such as the number of initial susceptibles and the initial number of infected in the population. Visualising the fitted parameters gives rise an isosurface plot of the feasible parameter ranges corresponding to each confidence level.

6.2 Future Work

We believe there are many areas of improvement for our project. In this section, we outline some ideas for potential extensions and further research:

- Estimate the starting time of the epidemic A potential extension to our methodology is to incorporate uncertainty inherent in the starting time of the epidemic. Currently, we assume this t_0 to be fixed, but in reality we do not know its true value. Traditionally, laborious manual work is being undertaken to detect the index case associated with an infectious disease spread in order to determine when did the epidemic emerge. Besides being very time consuming, such methods are also individuals being able to recall and provide complete, accurate information regarding their personal relationships.

- Implement uncertainty characterisation methodologies for other compartmental models So far, we have only focused on the SIR model, which is the most simple of the compartmental models. Another potential extension would be to develop analog methodologies for other, more complex compartmental models. The ability to characterise uncertainty in more realistic models would add great value to model-based analysis for policy and decision making.

- Develop a simulation-based methodology It is expected that real systems are likely to exhibit different characteristics than the ideal ones assumed by the classical SIR model; for example, real systems may feature time-varying parameters and the homogeneous mixing assumption may not apply. Nevertheless, the models may have utility in predicting the stochastic impact of candidate interventions in real systems with bounds [30], and a simulation-based methodology for this will could be a starting point for future work.

- Investigate model selection methodologies Development of complex epidemiological models increased the popularity of large scale simulations of epidemic spread in the literature. Accurate predicting on how an infection may spread is limited by the lack of rigorous approaches to validate such models and assess which one would be best for a particular problem. Furthermore, if we encounter a high goodness-of-fit for a set of observed data, how can we infer which specific model has produced it? Such questions do not have an answer yet and could be addressed in future work.

Bibliography

[1] A.J. Tatem, D.J. Rogers, and S.I. Hay. Global transport networks and infectious disease spread. In *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*, volume 62 of *Advances in Parasitology*, pages 293 – 343. Academic Press, 2006. Available online at http://www.sciencedirect.com/science/article/pii/S0065308X0562009X.

[2] S. Scott and C. Duncan. *Return of the Black Death. The World's Greatest Serial Killer*. Wiley, 2004.

[3] T.D. Hollingsworth, N.M. Ferguson, and R.M. Anderson. Frequent travelers and rate of spread of epidemics. *Emerging Infectious Diseases*, 19(9). Available online at http://wwwnc.cdc.gov/eid/article/13/9/07-0081.htm.

[4] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little Brown, 2000.

[5] J. Cannarella and J.A. Spechler. Epidemiological modeling of online social network dynamics. Available online at http://arxiv.org/abs/1401.4208, January 2014.

[6] Facebook statistics, 2014. Available online at http://www.statisticbrain.com/facebook-statistics/. Accessed on January 30, 2014.

[7] R.M. Christley, M. Mort, B. Wynne, J.M. Wastling, A.L. Heathwaite, R. Pickup, Z. Austin, and Latham S.M. "wrong, but useful": Negotiating uncertainty in infectious disease modelling. *PLoS NE*, 8(10), 10 2013. Available online at http://dx.doi.org/10.1371%2Fjournal.pone.0076277.

[8] M. Nika, G. Ivanova, and W.J. Knottenbelt. On celebrity, epidemiology and the internet. In *Proc. 7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, Turin, Italy, December 2013.

[9] Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

[10] M.J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

[11] M.J. Keeling. Models of foot-and-mouth disease. *Proceedings of the Royal Society, Biological Sciences*, 272(1569). Available online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1564112/?report=classic.

[12] Emilia Vynnycky and Richard White. *An introduction to infectious disease modelling*. Oxford University Press, 2010.

[13] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772): 700–721, 1927. Available online at http://rspa.royalsocietypublishing.org/content/115/772/700.short.

[14] H.W. Hethcote. The mathematics of infectious diseases. *SIAM Rev.*, 42(4): 599–653, December 2000. Available online at http://dx.doi.org/10.1137/S0036144500371907.

[15] J. M. Heffernan, R. J. Smith, and L. M. Wahl. Perspectives on the basic reproductive ratio. *Journal of the Royal Society, Interface / the Royal Society*, 2(4): 281–293, September 2005. Available online at http://dx.doi.org/10.1098/rsif.2005.0042.

[16] J. Lloyd-Smith. Parameter estimation, uncertainty, model fitting, model selection, and sensitivity and uncertainty analysis, June 2007. Lecture notes for the DIMACS/SACEMA/AIMS Advanced Study Institute, available online at dimacs.rutgers.edu/Workshops/AIMS/slides/JLS_ASI_6.pdf.

[17] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[18] A.F. Burns and W.C. Mitchell. *Measuring Business Cycles*. National Bureau of Economic Research, Inc, 1946. Available online at http://EconPapers.repec.org/RePEc:nbr:nberbk:burn46-1.

[19] M. Aoki. *New Approaches to Macroeconomic Modeling: Evolutionary Stochastic Dynamics, Multiple Equilibria, and Externalities as Field Effects*. Cambridge University Press, June 1996.

[20] F. Tramontana and M. Gallegati. Economics as a compartmental system: a simple macroeconomic example. Working Papers 1011, University of Urbino Carlo Bo, Department of Economics, 2010. Available online at http://EconPapers.repec.org/RePEc:urb:wpaper:10_11.

[21] Bob Duffy, Kevin Regan, Steve Coulombe, and John Yozzo. A real cliffhanger. 2012 retail report. Technical report, FTI Consulting, 2012. Available online at http://origin.fticonsulting.co/global2/media/collateral/united-states/2012-annual-retail-report-a-real-cliffhanger.pdf.

[22] C.C. Zou, W. Gong, and D. Towsley. Code red worm propagation modeling and analysis. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, CCS 2002, pages 138–147. Available online at http://doi.acm.org/10.1145/586110.586130.

[23] L.X. Yang and X. Yang. A new epidemic model of computer viruses. *Communications in Nonlinear Science and Numerical Simulation*, 19(6):1935 – 1944, 2014. Available online at http://www.sciencedirect.com/science/article/pii/S1007570413004656.

[24] J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright. Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998. Available online at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.6062.

[25] W.D. Flanders and D.G. Kleinbaum. Basic models for disease occurrence in epidemiology. *International Journal of Epidemiology*, 24(1):1–7, 1995. Available online at http://ije.oxfordjournals.org/content/24/1/1.abstract.

[26] B. Bolker and S. Ellner. Likelihood and all that, for disease ecologists. Tutorial distributed by The King Laboratory of Theoretical Ecology & Evolution at the University of Michigan, June 2011. Available online at http://kinglab.eeb.lsa.umich.edu/EEID/eeid/ecology/mle_2012.pdf.

[27] R. Dolgoarshinnykh. Introduction to epidemic modelling. Lecture Notes for a Graduate Course in Epidemic Modelling. Department of Statistics, Columbia University, 2005.

[28] T. Seidenfeld. *Philosophical problems of statistical inference: Learning from RA Fisher*. Number 22. Springer, 1979.

[29] D. Patterson. Profile likelihood confidence intervals for glm's. Lecture Notes for Mathematical Statistics Course, Department of Mathematical Sciences, University of Montana, 2011. Available online at www.math.umt.edu/patterson/ProfileLikelihoodCI.pdf.

[30] T. L. Burr and G. Chowell. Observation and model error effects on parameter estimates in susceptible-infected-recovered epidemiological models. *Far East Journal of Theoretical Statistics*, 19(2):163–183, 2013.

[31] F. Brauer. Compartmental models in epidemiology. In *Mathematical Epidemiology*, volume 1945 of *Lecture Notes in Mathematics*, pages 19–79. Springer Berlin Heidelberg, 2008. Available online at http://dx.doi.org/10.1007/978-3-540-78911-6_2.

[32] L. M. A. Bettencourt, A. Cintron-Arias, D. I. Kaiser, and C. Castillo-Chavez. The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models. 2005. doi: 10.2172/990668.