Imperial College
London

MSci Project

Imperial College London

Department of Physics

Epidemiological Modelling Of COVID-19

Author CID:
01519545

Project Code:
THEO-Contaldi-2

Word Count:
9985

Supervisor:
Prof. Carlo R. Contaldi

Assessor:
Prof. Gavin J Davies

Modelling the COVID-19 pandemic

As of the 1^{st} of March 2022, the global pandemic due to COVID-19 has resulted in 6,151,270 deaths across the globe. A truly global humanitarian crisis that has had substantial repercussions on almost everything that is connected to the globalised world we live in. As individuals, we have had to tolerate minor inconveniences, such as face masks, to complete transformations in the way we interact with the world and those around us. Amongst the torrent of change, we have all asked ourselves, why have we not stopped the pandemic already?

Governments and health organisation across the world have employed numerous measures such as face masks, social distancing, national lockdowns, and vaccines, with the aim of curbing the mounting infection cases. However, as we all know the success of these measures have varied, and it is all due to the extremely complex nature of our world. Before measures can even be implemented, governments need to understand how the pandemic will evolve, doing so will provide critical insight into what measures will be appropriate. Research institutions around the world have employed state-of-the art computational models to understand and predict the pandemic. Some have been highly effective whilst others have not. However, these models are generally resource and computationally expensive, requiring teams of scientists and a host of computational power to develop them.

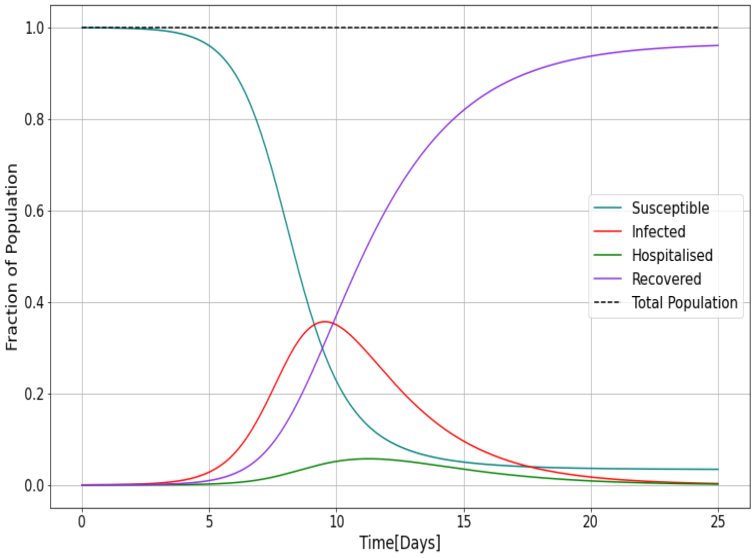

Figure 1: Evolution of the SIHR compartmental model

This is where our MSci project comes in, we use relatively simple and computationally efficient deterministic compartmental models to understand the pandemic. Deterministic compartmental models, split a population of individuals into compartments, where each compartment contains the individuals in a population that are in a specific state. For example, the susceptible, infected, and recovered (SIR) model splits the population into those that can get infected, are currently infected and those that have recovered, respectively. The flow of individuals between compartments are described by a system of differential equations, which when we solve, shows us how the number of individuals in each compartment change over the period of the pandemic. The equations of these compartmental models include variables known as parameters. These essentially govern how the pandemic will evolve, tell us how infectious a disease is and how quickly individuals recover. These parameters are vital not only because they let us model the pandemic, but also, they provide critical information on the disease itself. A general plot of the SIHR (susceptible, infected, hospitalised and recovered) compartmental model is shown in figure (1), showing how each compartment in the model evolves over a period. In our project we use the SIR and the SIHR compartmental models to model the evolution of the COVID-19 pandemic in England for the period, July 2020 to January 2021. We use sampling methods to find the set of parameters that develop a model that most closely represents the pandemic data. We show that these simple models can predict the general trend of the data for up to three weeks, after which the parameters need to be updated.

Contents

Abstract

This report details an investigation into the SIR (susceptible, infected and recovered), SIHR (susceptible, infected, hospitalised and recovered) and the SVEASyHRD (susceptible, vaccinated, exposed, asymptomatic, symptomatic, hospitalised, recovered and deceased) compartmental models and their use in modelling epidemics. The report then discusses the procedure of using Markov Chain Monte Carlo (MCMC) sampling methods to find the parameters of the SIR and SIHR models that best fit to simulated data of an epidemic. The report then discusses the use of the SIR and SIHR compartmental models in conjunction with MCMC to find the parameters of the COVID-19 pandemic for the period July 2020 to January 2021. The report investigates the use of constant and time-varying parameters in modelling the pandemic. Time-varying parameters are shown to be superior for fitting and predicting the evolution of the COVID-19 epidemic. We use the parameters we determined, $\beta = 2.5348 \pm 0.00023$ and $\gamma = 2.50422 \pm 0.0001$, found for the SIR model and the parameters, $\beta = 2.474 \pm 0.051$, $\gamma = 2.396 \pm 0.068$, $\mu = 0.041 \pm 0.031$ and $\epsilon = 0.023 \pm 0.033$, for the SIHR model, to predict the future cases of the pandemic for the period 06/11/2020 to 31/12/2020. We show that the models can predict the general trend from 06/11/2020 to 07/12/2020, however after this point the epidemic evolves and the determined parameters no longer apply.

Chapter 1

Introduction

The current global health crisis is centred around the outbreak of coronavirus disease 2019 (COVID-19), a contagious respiratory disease that can lead to pneumonia, hypoxia, respiratory failure, and death. The cause of the disease is a novel coronavirus that causes severe acute respiratory syndrome (SARS) [1]. The outbreak of SARS-CoV-2 (coronavirus disease 2019) initially started due to its transmission from animals to humans in the Huanan seafood market in Wuhan, China, and has subsequently spread to 221 countries and mutated into different variants [2]. In January 2020, the WHO declared a global health emergency due to the exponentially growing rates of infection in China and other countries [2]. Like other instances of SARS (2002 - 2003) and the Middle East respiratory syndrome (MERS) (2012 - Present), the coronaviruses can infect humans and a large range of animals through the air [2, 3]. COVID-19 has had substantial repercussions on social, economic, and public healthcare sectors as governments and health organisations attempt to reduce infection and mortality rates.

The pandemic has severely impacted the U.S. and world economy due to the implementation of national lockdowns and other preventative measures. Between January and July 2020, the U.S. unemployment rate rose from 3.6 % to 10.1 % and 115 million Americans experienced a loss in employment income. The U.S. stock prices fell logarithmically by 42% between February and March 2020 until the Federal Reserve put special measures into place [4, 5]. As of the 1^{st} of March 2022, the number of confirmed deaths due to the virus is 6,151,270 with 490,672,237 cases across the globe [6]. To reduce mounting infection rates countries such as the UK imposed preventative measures such as face masks, social distancing, self-isolation, national lockdowns, and vaccines [7].

However, for government agencies to make informed decisions, they need to understand how an epidemic will evolve. Mathematical models are used to describe and predict how an infectious disease will propagate through a population. These models are also utilised to find parameters of an epidemic, such as how transmissible a disease is, to calculate the effect of strategies such as lockdowns or vaccinations [8]. During the pandemic, research institutions utilised their expertise in epidemiological modelling to create different mathematical models with the aim of successfully understanding how the UK population would be affected. CovidSim for example, an agent-based model (ABM) developed by the Imperial College COVID-19 Response Team was an instrumental tool in informing the UK government in changing its policy and implementing a lockdown to limit the spread of the Coronavirus [10].

Epidemic models can be put into two groups, stochastic and deterministic. Stochastic models estimate the potential outcomes of a disease by randomly varying inputs over time. These models aim to capture the inherent randomness in disease transmission and are most effective when describing emerging epidemics where case numbers are small [8,9]. Deterministic models, also known as compartmental models, divide a population into subgroups or compartments, each representing a specific state. Compartmental models emerged in the 1920s through the Kermack-McKendrick epidemic model. This specific model described the relationship between susceptible, infected, and immune (recovered) individuals in a population where the transition rates of each compartment were described by a system of differential equations [8, 11].

Compartmental models can be further developed by utilising them in conjunction with Agent-based models as done by CovidSim. Agent-based models are computational models that simulate the interactions of individuals to predict how a system will evolve. Models such as CovidSim have been shown to be more accurate when compared to just compartmental models, however, they are also significantly more demanding in terms of computational power and resources due to having vast amounts of features [10]. The aim of this project is to utilise relatively simple models, in comparison to those used by research institutions to find and understand the parameters of the COVID-19 pandemic and to be able to predict the evolution of the pandemic using real data.

One of the main objectives of this project was to investigate compartmental models with varying complexities to understand how the dynamics of an epidemic change. In this report, we investigate the SIR (susceptible, infected and recovered), SIHR (susceptible, infected, hospitalised and recovered) and the SVEASyHRD (susceptible, vaccinated, exposed, asymptomatic, symptomatic, hospitalised, recovered and deceased) compartmental models. The second objective of the project was to build and use Markov Chain Monte Carlo sampling methods to find the parameters of the SIR and SIHR compartmental models that would best describe simulated data. We then evaluate the accuracies of the fitted models and understand the distribution of the parameters. The final objective of the project was to use the SIR and SIHR compartmental models and Markov Chain Monte Carlo methods and apply them to real data of the COVID-19 pandemic. Due to COVID-19 evolving over the course of the pandemic, including the emergence of different variants, in this report, we investigate the period of July 2020 to January 2021 and utilise data directly from the UK government. We then use the fitted models to find the parameters of the pandemic and use these to predict how the pandemic will evolve.

Chapter 2

Compartmental Models

Developing and understanding different compartmental models forms a significant portion of this project. Compartmental models assign individuals in a population to a specific compartment and use ordinary differential equations (ODE) to describe the rate of change in each compartment [11].

In this project we explore the SIR model and use the mathematical foundation to develop our own models which include the susceptible, infected, hospitalised and recovered (SIHR) model and the susceptible, vaccinated, exposed, asymptomatic, symptomatic, hospitalised, recovered and deceased (SVEASyHRD) model. Each model's system of ODE needs to be solved to obtain the number of individuals in each compartment at each time step. Compartmental models do not have analytical solutions. Instead, we treat the ODE as an initial value problem, where we specify the initial conditions of the system and utilise ODE solvers to obtain a numerical solution. In this chapter we describe the mathematical development of each of the models, the methodology to find the population of each compartment and the results, showing how the compartments in each model evolve.

2.1 Theory

Compartmental models make several assumptions to simplify the complexity of real epidemics. Firstly, we assume that there is no prior immunity, which means that all individuals in a population have the same potential to become sick. We assume all individuals are the same in terms of spreading and contracting the disease. We assume homogeneous mixing of the population, meaning the probability of encountering other individuals in a population is the same. We also assume that there are no behavioural changes during the epidemic [11].

2.1.1 SIR Compartmental Model

The SIR model is one of the most basic compartmental models. Susceptible, S, means the number of individuals that can become infectious and transition into the infected compartment. Infected, I, means the number of people than can infect others with the disease. Recovered, R, means the number of individuals that can no longer contract the disease as they have recovered from being infected [11]. We study the SIR model without vital dynamics. This means we ignore natural births and deaths within a population, assuming a constant population size [12, 13].

The parameters of the SIR model, β and γ, shown in figure (2.1) govern the magnitude of the rate of transition between their respective compartments. The flowchart shows the interaction

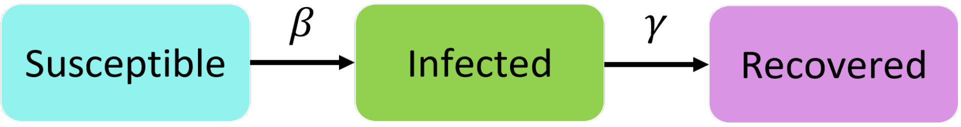

Figure 2.1: Flowchart showing the susceptible (S), infected (I) and recovered (R) compartments of the SIR compartmental model. The parameter, β, is the average number of contacts per person per time between susceptible and infectious individuals. The parameter, γ, is the probability an infected individual will no longer remain infected. The parameters of the model control the magnitude of the transition rates between compartments. Arrows indicate the direction of transition between compartments.

between each compartment and the direction of flow (arrows) which can then be used to obtain the differential equations for the system. The system of ODE for the SIR compartmental model are shown below,

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \qquad \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \qquad \frac{dR}{dt} = \gamma I, \tag{2.1}$$

where $\frac{dS}{dt}$, $\frac{dI}{dt}$ and $\frac{dR}{dt}$ are the rate of change of the S, I and R compartments respectively and N is the total population size given by N = S + I + R. The parameter β, is the average number of contacts per person per time [14]. $\frac{SI}{N}$ gives us the probability of disease transmission in a contact between a susceptible and an infectious individual and thus multiplying the term by β gives us the number susceptible individuals who become infected. The parameter, γ, is the probability of a contagious person becoming non-contagious and conversely $\frac{1}{\gamma}$ is the average number of days an infectious person remains infectious. γI tells us the number of individuals who recover from the disease. The values of S, I and R are functions of time measured in number of days [14].

The parameters in equation (2.1) allow us to derive another relation,

$$R_0 = \frac{\beta}{\gamma}, \tag{2.2}$$

where R_0 is the basic reproduction number. The R_0 of an infection is the expected number of people that will be infected from a single infected individual [15]. The reproduction number is important in determining whether an infectious disease will spread through a population. An R_0 >1 means that the infection will spread through a population with greater values signifying increased difficulty in controlling the epidemic [16].

2.1.2 SIHR Compartmental Model

The SIHR model introduces the hospitalised (H) compartment. H means the number of individuals that are infected and in the hospital, due to rapidly failing health. Like the SIR model we assume that the population size is constant.

Figure (2.2) shows the flowchart for the SIHR model. In comparison to the SIR model, the SIHR model introduces two new parameters μ and ϵ which are the probability an individual will transition from being infected to being hospitalised and from being hospitalised to recovered, respectively. Using the flowchart, we see that, instead of all infected individuals eventually entering the recovered compartment, a portion of those will first enter the hospitalised compartment and then the recovered compartment. Introducing the new parameters and compartment to equation (2.1), the system of ODE for the SIHR compartmental model are shown below,

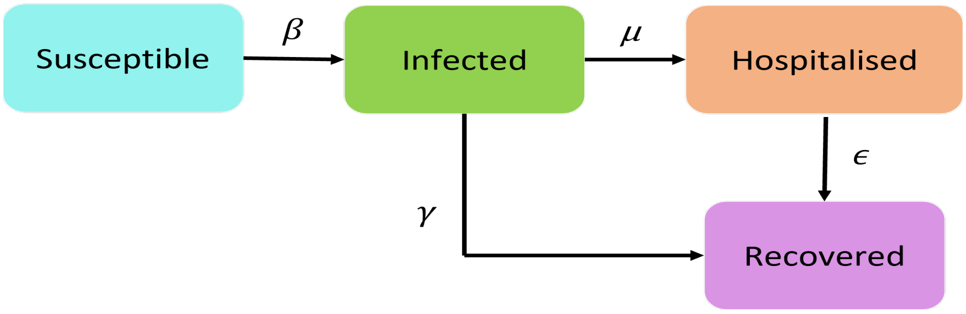

Figure 2.2: Flowchart showing the susceptible (S), infected (I), hospitalised (H) and recovered (R) compartments of the SIHR compartmental model. The parameter, β, is the average number of contacts per person per time between susceptible and infectious individuals. The Parameters μ, γ and ϵ are the probability an individual will transition from being infected to hospitalised, infected to recovered and hospitalised to recovered respectively. Parameters of the model control the magnitude of the transition rates between their respective compartments. Arrows indicate the direction of transition between compartments.

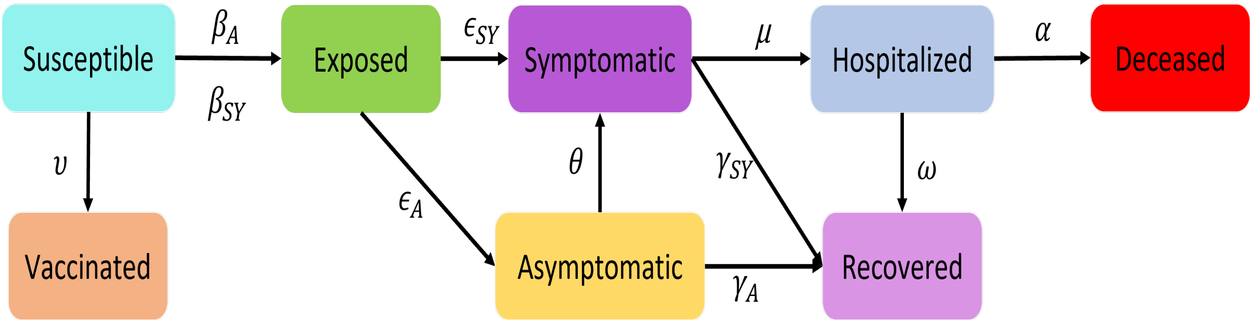

Figure 2.3: Flowchart showing the susceptible (S), vaccinated (V), exposed (E), asymptomatic (A), symptomatic (Sy), hospitalised (H), recovered (R) and deceased (D) compartments of the SVEASyHRD compartmental model. The parameters, β_A and β_{Sy}, are the average number of contacts per person per time between susceptible and asymptomatic or susceptible and symptomatic individuals respectively. The parameters υ, ε_{Sy}, ε_A, θ, γ_{Sy}, γ_A, μ, ω and α are the probability an individual will transition from their originating compartment to their destination compartment, indicated by the direction of the arrows. The parameters of the model control the magnitude of the transition rates between their respective compartments.

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \qquad \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I - \mu I \qquad \frac{dH}{dt} = \mu I - \epsilon H \qquad \frac{dR}{dt} = \gamma I + \epsilon H, \tag{2.3}$$

where the differential $\frac{dH}{dt}$ is the rate of change of the hospitalised compartment and the total population is defined as N = S + I + H + R.

2.1.3 SVEASyHRD Compartmental Model

This report aims to investigate the use of relatively simple compartmental models. However, here we now discuss the SVEASyHRD model. We introduce the exposed compartment, which means the number of individuals who have been exposed to the disease but are currently not infectious. The asymptomatic compartment, for individuals who have the disease but do not show symptoms and the symptomatic compartment, for individuals who show symptoms. We also introduce the deceased compartment, for individuals who have died due to the virus and the vaccinated compartment, for vaccinated individuals. In this model individuals die and therefore must be removed from the total population.

Figure (2.3) shows the flowchart for the SVEASyHRD model. The parameters β_A and β_{Sy}, are the average number of contacts per person per time between susceptible and asymptomatic or susceptible and symptomatic individuals respectively. The parameters υ, ε_{Sy}, ε_A, θ, γ_{Sy}, γ_A, μ, ω and α are the probability an individual will transition from their originating compartment to their destination compartment, shown by the direction of the arrows. In comparison to the SIR and SIHR, this model introduces a few new compartments and parameters. Like the previous models, the parameters control the magnitude of the transition rates. The system of ODE for the SVEASyHRD model are shown below;,

$$\begin{aligned}
&\frac{dS}{dt} = -\frac{S}{N}(A\beta_A + Sy\beta_{Sy}) - \upsilon S \quad && \frac{dE}{dt} = \frac{S}{N}(A\beta_A + Sy\beta_{Sy}) - E(\varepsilon_A + \varepsilon_{Sy}) \quad && \frac{dV}{dt} = \upsilon S \\
&\frac{dA}{dt} = \varepsilon_A E - \gamma_A A - \theta A && \frac{dSy}{dt} = \varepsilon_{Sy} E - \gamma_{Sy} Sy - \mu Sy + \theta A && \frac{dD}{dt} = \alpha H \\
&\frac{dR}{dt} = \gamma_A A + \gamma_{Sy} Sy + \omega H && \frac{dH}{dt} = \mu Sy - \alpha H - \omega H, &&
\end{aligned} \tag{2.4}$$

where the differential equations $\frac{dV}{dt}$, $\frac{dE}{dt}$, $\frac{dA}{dt}$, $\frac{dSy}{dt}$ and $\frac{dD}{dt}$ are the rate of change of the vaccinated, exposed, asymptomatic, symptomatic and deceased compartments respectively. The total population of the system, is defined as N = S + E + V + A + Sy + H + R.

2.2 Methodology

In this section, we discuss how we solved the models developed in section 2.1. Analytical solutions do not exist hence we obtain numerical solutions to these models, treating them as initial value problems to obtain a value for each compartment at each time step. To do this we utilise the Python programming language and the solve_ivp function from the Scipy Python package [17]. This function numerically integrates a system of ODE given initial values. The methodology of obtaining the numerical solutions is shown below.

1. Define model
   - Create functions, each containing a system of ODE as described by equations (2.1), (2.3) and (2.4).
   - These functions require: parameters of the model, total population (N), array of time values and an array that contains the values of each compartment in the previous iteration.
   - The function then computes and returns the value of each differential equation for the current iteration.
2. Define initial conditions
   - Create an array with the initial number of individuals in each compartment.
   - Create an array with the values of the parameters used in the model including N.
   - Create an array of time values over which we want to evolve our epidemic.
3. Utilise solve_ivp
   - We provide the integrator: the function counting the compartmental model, array of time values, array of initial compartment values and the array containing the parameters.

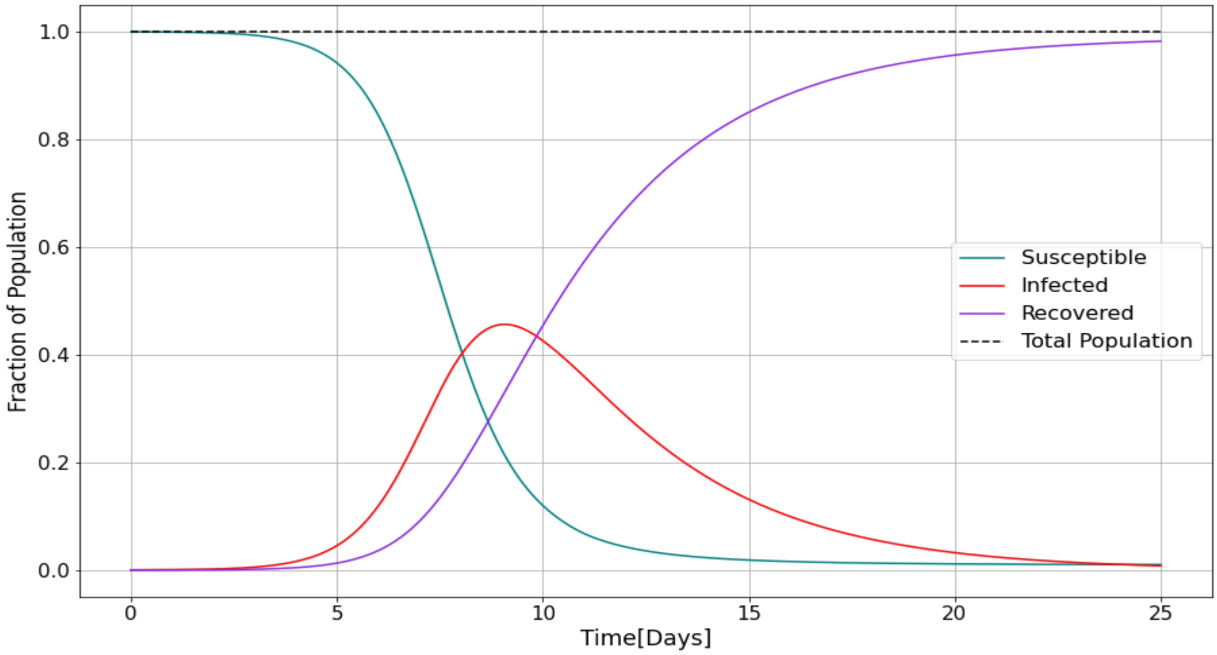

Figure 2.4: SIR compartmental plot, showing the fraction of the population in each compartment as a function of time. The evolution of the compartments is described by equation (2.1). The parameters used in the model are $\beta = 1.4$ and $\gamma = 0.3$. The initial values of the S, I and R compartments are 9980, 2 and 0 respectively. The total population line, monitors the sum of all compartments.

- The function then integrates and returns an array of values for each compartment. Each value in the array represents the number of individuals in that compartment at the corresponding time point.

Using the methodology listed above we then plot the array of values for each compartment as a function of time showing how each compartment evolves in relation to one another during the epidemic.

Implementing compartmental models requires several assumptions at this stage. Firstly, we use random values for the model parameters. This is because we do not want to make any initial assumptions about the parameters, and we want to gain a general understanding of the model's behaviour. We measure time in number of days as this is most appropriate when considering the frequency of new data. We utilise an arbitrary but large value for N as required for deterministic models to function correctly.

To verify that the models were implemented correctly, we tracked the total population size through the integrator. The models described in figures (2.1) and (2.2) should have a constant N. We also checked the implementation by ensuring at each iteration the sum of the differentials for the models is zero.

2.3 Results and Discussion

In this section we discuss the numerical solutions of the compartmental models.

2.3.1 SIR Compartmental Model

We integrate equation (2.1) with parameters $\beta = 1.4$ and $\gamma = 0.3$ with initial values of the S, I and R compartments as 9980, 2 and 0 respectively.

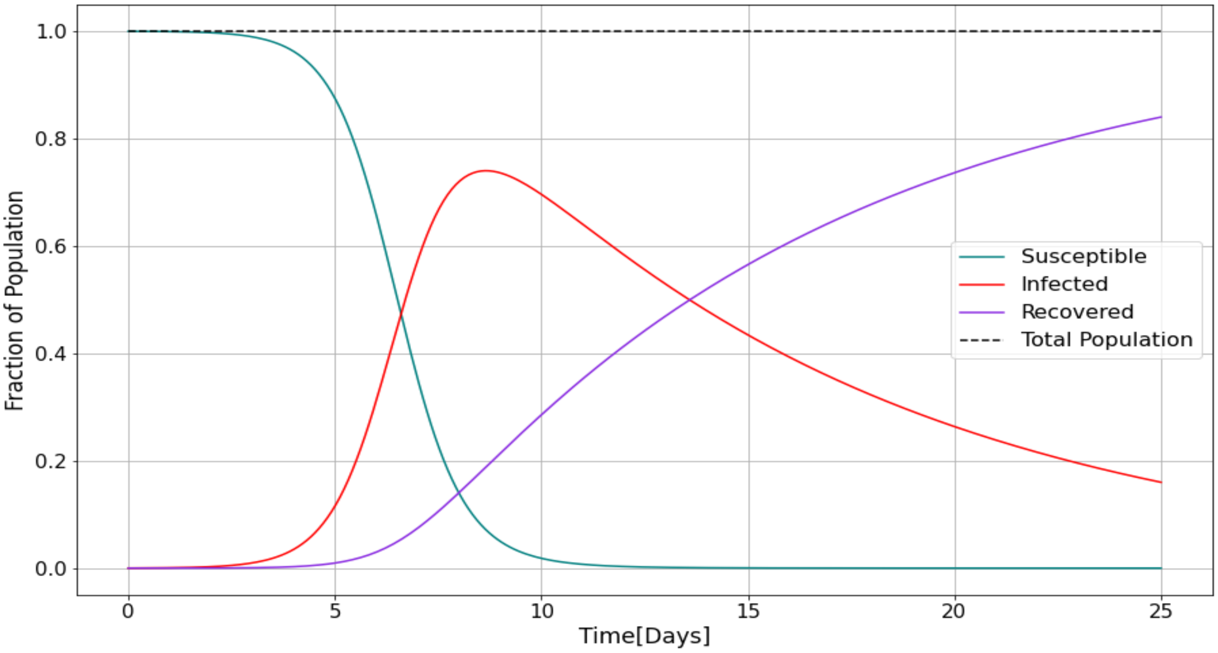

Figure 2.5: SIR compartmental plot, showing the fraction of the population in each compartment as a function of time. The evolution of the compartments is described by equation (2.1). The parameters used in the model are $\beta = 1.4$ and $\gamma = 0.1$. The initial values of the S, I and R compartments are 9980, 2 and 0 respectively. The total population line, monitors the sum of all compartments.

The numerical solution is shown in figure (2.4). We plot how the population of each compartment, measured as a fraction of the total population, changes as a function of time. The figure shows that initially there is little change in the number of susceptible individuals, which is as expected as our β is small and the initial number of infected individuals is only 0.02% of the population. As more individuals get infected, we see an exponential decrease in the number of individuals in the susceptible compartment. At the same time the number of infected individuals rises till it peaks at 45% of the population and then approaches zero. The number of recovered individuals initially increases exponentially as more infected individuals transition to the compartment, but as the infected population decreases it begins to plateau. The infected compartment has a defined peak because of $\gamma = 0.3$ which means the average number of days an individual is infected is, $\frac{1}{\gamma} = 3.3$, which is a relatively short time period. The implementation of the model is verified by the total population line which remains constant at 1.0 throughout the epidemic.

In figure (2.5) we utilise the same model and initial conditions as figure (2.4) but change the parameter, γ. to $\gamma = 0.1$. The figure shows that initially the susceptible compartment has minimal change, like figure (2.4) which is expected as β is the same. We notice however that the maximum number of infected individuals rises to 75% of the total population and the number of infected individuals decreases at a slower rate than in figure (2.5). This result is expected because here $\frac{1}{\gamma} = 10$, meaning that individuals remain on average, infectious for 10 days, 7 days more than figure (2.4).

2.3.2 SIHR Compartmental Model

We integrate equation (2.3) with parameters $\beta = 1.4$, $\gamma = 0.3$, $\mu = 0.1$ and $\epsilon = 0.5$ with initial values of the S, I, H and R compartments as 9980, 2, 0 and 0 respectively.

Figure (2.6) shows the numerical solution to the SIHR model. A key difference between figure

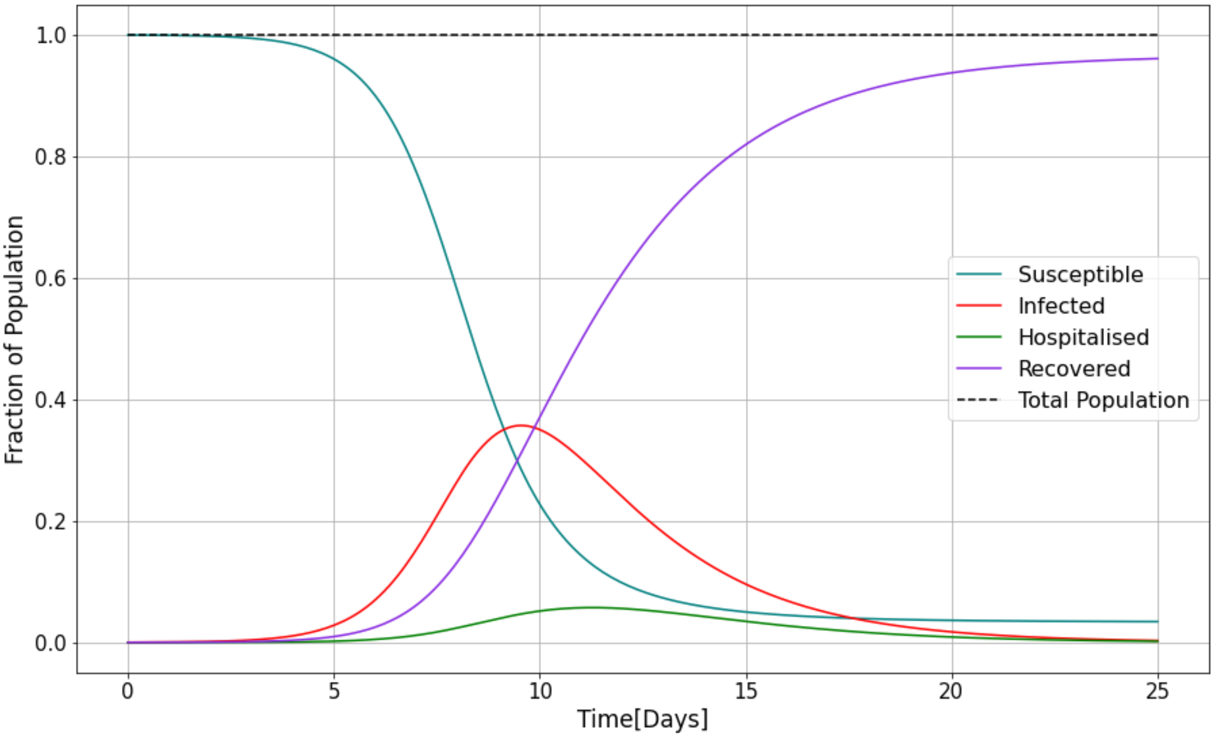

Figure 2.6: SIHR compartmental plot, showing the fraction of the population in each compartment as a function of time. The evolution of the compartments is described by equation (2.3). The parameters used in the model are $\beta = 1.4$, $\gamma = 0.3$, $\mu = 0.1$ and $\epsilon = 0.5$. The initial values of the S, I, H and R compartment as 9980, 2, 0 and 0 respectively. The total population line, monitors the sum of all compartments.

(2.4) and (2.6) is that we introduce the hospitalised compartment. As a result, the maximum number of infected individuals at a given time decreases to 36%. This is expected as individuals can now transition from being infected to hospitalised. We see that the number of individuals that are hospitalised is a relatively small fraction of the total population, and this is expected as $\mu < \epsilon$.

2.3.3 SVEASyHRD Compartmental Model

We integrate equation (2.4) with initial compartment values shown in table (2.1) and parameter values shown in table (2.2).

Compartment	S	V	E	A	Sy	H	R	D
Value	9980	0	15	10	5	0	0	0

Table 2.1: Table showing the initial compartment values for the SVEASyHRD model.

Parameter	β_A	β_{Sy}	υ	ε_{Sy}	ε_A	θ	γ_{Sy}	γ_A	μ	ω	α
Value	0.75	0.4	0.01	0.5	0.25	0.003	0.1	0.25	0.15	0.09	0.06

Table 2.2: Table showing the parameter values for the SVEASyHRD model.

Figure (2.7) shows the numerical solution to the SVEASyHRD model. As discussed in this chapter, models can be made to contain more compartments than the basic SIR model. We could extend these models by including an age or social distancing compartment. Introducing more compartments can make these models more realistic and reflective of the true nature of an

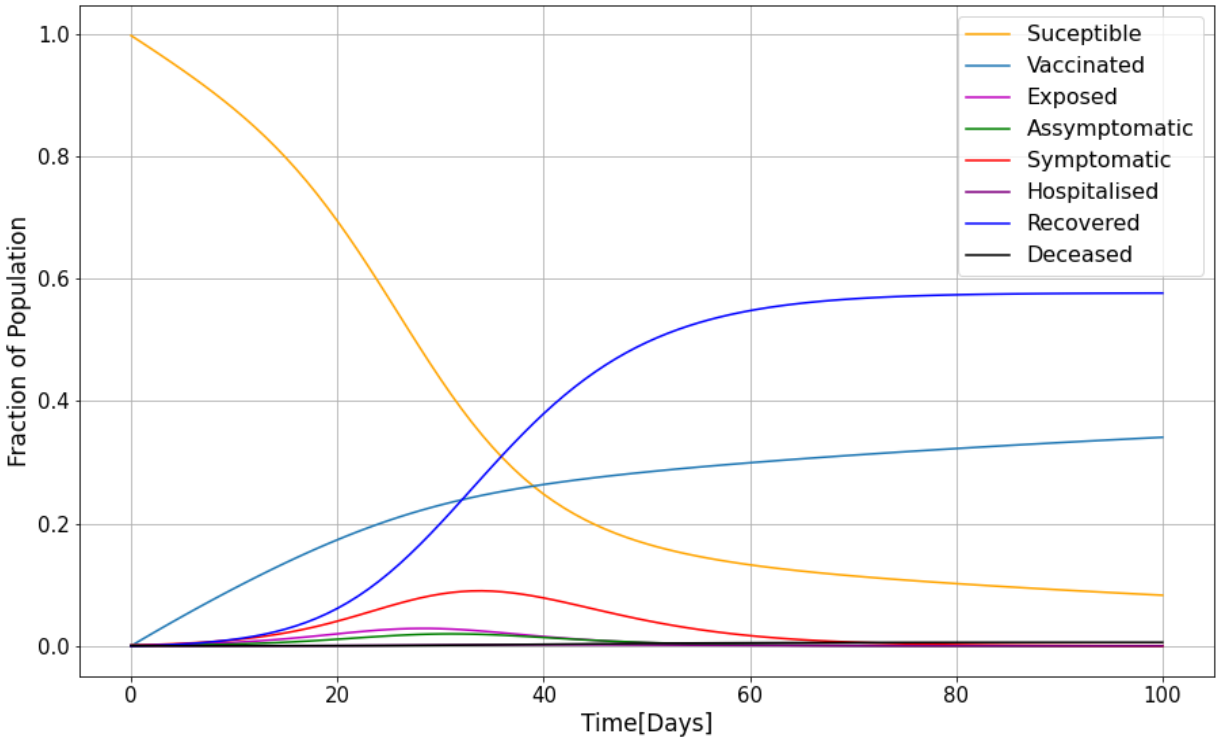

Figure 2.7: SVEASyHRD compartmental plot, showing the fraction of the population in each compartment as a function of time. The evolution of the compartments is described by equation (2.4). The parameters used in the model are shown in table (2.2). Initial values of the compartments are shown in table (2.1).

epidemic. However, it is impossible to know how a population will be partitioned. As we make models more complex, the more prone the model becomes to small errors in the estimation of each parameter. Due to compartmental models being exponential systems, these errors can propagate through the system and increase. As a result, in this project we build compartmental models that are descriptive enough (SIR and SIHR), with only a few parameters whose correct statistical treatment we can ensure.

Chapter 3

Investigating Simulated Data

A core objective of this project was to be able to fit a compartmental model to simulated data and find the parameters of this fitted model. In this project we use Markov Chain Monte Carlo (MCMC) sampling to find the optimal parameters of a model. In this chapter we describe the theoretical principle behind MCMC and the methodology we use to develop our own algorithm. We then discuss the implementation of PYMC3, a Python package that performs MCMC sampling, and discuss how we developed simulated data.

3.1 Theory

3.1.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a class of algorithms for sampling from a probability distribution [18]. It combines two properties: Monte-Carlo and Markov chain. Monte-Carlo is the procedure of numerically estimating the properties of a distribution by taking random samples from the distribution itself [19]. The Markov chain is where each random sample in the process is used to generate the next random sample, forming a chain of random samples [19]. The goal of employing MCMC sampling here is to find parameters that develop a model that most closely represents the data.

Bayesian inference forms the foundation of MCMC sampling [19, 20]. It is a statistical inference method based on Bayes' formula,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{3.1}$$

where θ is the parameters and $P(\theta)$ is the prior probability before the data D is observed. $P(\theta|D)$ is the probability of the parameters given the data, known as the posterior probability. $P(D|\theta)$ also known as the likelihood is the probability of observing D given θ. $P(D)$ is known as the marginal likelihood. The underlying principle behind Bayes' equation is to update our old hypotheses (priors) to obtain new hypotheses (posteriors) [21]. The posterior in turn allows us to determine the probability distribution of each parameter in a model and to infer the optimal parameter value and uncertainty.

The marginal likelihood, $P(D)$, is essentially a normalising constant that is not required when considering different parameters for a single model, this allows us to rewrite Bayes' equation as,

$$P(\theta|D) = P(D|\theta)P(\theta) \tag{3.2}$$

where the prior $P(\theta)$ can be set to 1 if we assume a uniform prior. Equation (3.2) is critical to MCMC sampling because it allows us to state that the posterior probability is equal to the likelihood assuming a uniform prior [22]. MCMC uses the concept behind updating the hypotheses to approximate the posterior probability distribution. This is done by obtaining $P(D|\theta)$ through an iterative procedure where we randomly sample sets of parameters to find the optimal set of parameters, essentially maximising the likelihood. This is equivalent to minimising the error between the model and the actual data. In doing this process we develop a chain of accepted values for each parameter. The histogram of the values of each parameter in the chain form the posterior distribution of that parameter [23].

MCMC can be used to determine the parameters according to a criterion defined as,

$$\chi^2 = \sum_i \chi_i^2 \tag{3.3}$$

where χ_i^2 is the measure of error of an individual compartments in the model. This is defined as,

$$\chi_i^2 = \frac{1}{2} \sum_j \frac{(D_i(t_j) - M_i(t_j|\theta))^2}{\sigma_i^2} \tag{3.4}$$

$D_i(t_j)$ is the data for a specific compartment, $M_i(t_j|\theta)$ is the corresponding compartment of a model with a set of parameters θ and σ is the standard deviation of the data [23]. The criterion χ^2 is used to accept or reject a certain set of parameters with the aim of minimising the χ^2 [24, 25].

3.1.2 Simulating Data

We simulated data using the principle,

$$D = S + n \tag{3.5}$$

where D is the simulated data, S is the underlying signal and n is noise. We pass equation (3.5) a signal with known parameters and change it by introducing noise [25]. The MCMC algorithm is then passed the data, D, with the aim of finding the parameters used to develop the signal.

3.2 Methodology

In this section, we discuss the development of simulated data and our MCMC sampling methods.

3.2.1 Generating Simulated Data

We simulate data using the principle behind equation (3.5), the methodology is shown below.

1. Define a model
    - Select a compartmental model we want to simulate data for.
    - Specify initial conditions and parameters.

2. Create a model with noisy parameters
    - Solve the system of ODE for a specific period and introduce noise to the parameters in each iteration.
    - We introduce Gaussian noise to the parameters each time the integration function passes over the system of ODE. We do this by selecting parameters from a Gaussian distribution, centred at the original parameter with a standard deviation that is a fraction of the original parameter value.
    - At the end of the integration we receive an array of values for each compartment.
3. Introduce noise to the values of the compartments
    - Introduce Gaussian noise to the array of values for each compartment, by varying the values by a specific standard deviation.
    - Treat each modified array as a separate data source.

3.2.2 Developing Markov Chain Monte Carlo

Here we discuss the development of our own MCMC algorithm based closely on the Metropolis-Hastings algorithm [26]. We then discuss how we utilised PYMC3, an MCMC package that is a more efficient MCMC sampler. The algorithm and procedure for our own MCMC sampler is shown below.

1. Initialise the system
    - Obtain simulated data for one or more compartments.
    - Specify initial condition of each compartment. Create time array with same length as the data.
2. Create initial model
    - Pick a compartmental model containing at least one compartment that is the same as the data.
    - Pick initial parameters randomly from a uniform distribution and solve the system of ODE.
3. Compute χ^2
    - Utilise equation (3.3) and compute the χ^2 by summing individual χ_i^2 of each compartment.
    - Save the initial χ^2 and parameters.
4. Iterate over the algorithm
    - Repeat process in step 2, however now we pick parameters from a Gaussian distribution centred at the previously saved parameters with a standard deviation that is 0.01 of the parameter value.
    - Repeat method in step 3 with the new parameters.
        - If the $\chi^2 < \chi^2_{prev}$, we update our saved parameters and saved χ^2 to the new values. We then repeat the process as described above until maximum number of iterations.

    - If $\chi^2 > \chi^2_{prev}$, we introduce the concept of acceptance percentage. Where 20% of the time we accept the parameters and the χ^2 that gives us a χ^2 higher than the previous iteration. Otherwise, 80% of the time we reject the values. We then repeat the process as described in step 4 until the maximum number of iterations.

5. Report values for chain
    - Once the maximum number of iterations are completed, we save each accepted set of parameters as a list which forms a chain of parameters. We also report the χ^2 at each accepted point.
    - Steps 1 to 4 create a single chain. We then repeat this process to create multiple chains.

The MCMC process aims to minimise χ^2 and have a list of parameters that forms a chain. In the algorithm, we introduced the concept of acceptance percentage. We do this to sample the parameter space and understand the distribution of the parameters instead of performing a simple stochastic gradient descent, that would only provide us with the optimal set of parameters [26]. We use multiple chains to sample a much wider parameter space and to increases the probability of finding the global minimum. We then analyse the chain that produces the smallest χ^2.

Chains are analysed using a package called GetDist, to obtain the posterior distributions of the parameters [27]. GetDist uses the chain produced through MCMC sampling and histograms the parameters, integrating them across the other parameters to obtain 1D and 2D marginalised plots. These plots then allow us to infer the posterior distributions of each parameter. Before providing the chains to GetDist we need to 'burn-in' a chain. Here we remove the first n samples from the chain, essentially removing the points in the parameter space that may be far away from the global minimum and thus unnecessary when it comes to understanding the distribution of the optimal parameters [23].

We applied our MCMC sampling method to the SIR model, however for more complex models the algorithm was inefficient due to the large number of parameters. As a result, we decided to use PYMC3, a Python package that focuses on advanced Markov Chain Monte Carlo algorithms [28]. It utilises adaptive learning, where chains run in parallel and can communicate with each other to approach the global minimum efficiently. However, PYMC3 is a black box package that uses the Theano class to perform computations and is only simple to use with basic functions. In our case where a system of ODE had to be integrated at each iteration, significant adjusting was required to make the package work [28]. Below shows the methodology of how we utilised PYMC3.

1. Define a likelihood function
    - PYMC3 requires a likelihood function that it aims to maximise.
    - In the function we solve the system of ODE for specific parameters chosen by PYMC3 and compute -χ^2 (likelihood) for the data and the corresponding compartments.
    - Return the value of the likelihood.
2. Create a class to make Theano objects compatible with NumPy
    - Convert parameters that are passed to the class as Theano objects into NumPy objects.
    - Provide parameters to the likelihood function mentioned in step 1.

- Return value of the likelihood as a Theano object.

3. Initialise PYMC3 sampling
    - Define a PYMC3 model.
    - Initialise parameters of a specific compartmental model using a uniform distribution (provided by PYMC3) and as a Theano object.
    - Pass PYMC3 the class described in step 2 and the parameters of the model.
4. Perform MCMC sampling
    - Provide PYMC3 the number of samples and chains.

PYMC3 then returns the chains containing the sets of parameters.

In this chapter, we simulated the data for the SIR and SIHR compartmental models. We then use our own developed MCMC algorithm on the SIR simulated data and verify that the MCMC algorithm is implemented correctly. We then explore the marginalised posteriors of the model parameters and compare them to the known parameters. We then utilise PYMC3 to find the marginalised posteriors of the SIHR compartmental model parameters and compare the differences that arise in finding the optimal parameters of models with different compartments.

3.3 Results and Discussion

In this section, we discuss the results of simulating compartmental model data and the application of the MCMC sampling methods to find the posterior distributions of the parameters.

3.3.1 Applying SIR Compartmental Model

Define an SIR model with the parameters $\beta = 1.4$ and $\gamma = 0.3$. Each compartment, susceptible, infected, and recovered is an individual signal, with initial population sizes of 9980, 2 and 0 respectively. We introduce noise as discussed in the methodology to create simulated data. However, during a real pandemic, we would have data only on the number of infected individuals. As a result, we only develop infected data with the aim of using only the data from this compartment in the MCMC process. The infected data and the underlying signal are shown in figure (3.1).

Gaussian noise was introduced to the original parameters with a mean at the signal parameters, θ, and a standard deviation of 0.1θ. Gaussian noise was also introduced to the final values of the signal Infected curve, I, with a mean at the original signal values and a standard deviation of 0.1I. Figure (3.1) shows that although noise is introduced, the data has a similar structure to the underlying signal. During the MCMC process we provide the same initial compartment values as used in developing the original signal. We produce 5 chains each running for 16,000 iterations.

Figure (3.2) illustrates that the developed MCMC algorithm was implemented correctly. It shows how the parameter γ changes as the MCMC algorithm probes the parameter space. The results shown are the accepted values for γ in the optimal chain. The figure shows that the value of γ increases and then decreases as it finds the global minimum. It then reaches a point where the value fluctuates about a constant point, which is the global minimum. The Burn-In line indicates the point before which samples need to be removed to analyse the distribution of the parameter space about the minimum.

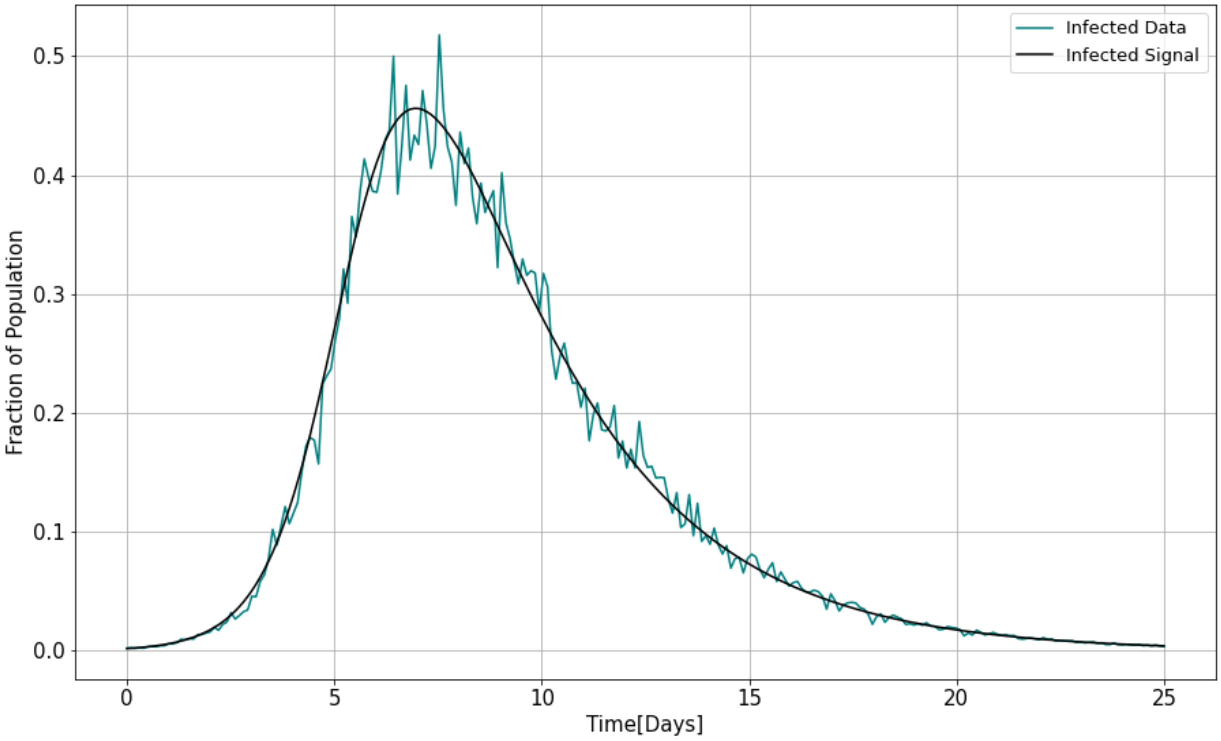

Figure 3.1: SIR compartmental model plot, showing the fraction of the population that is infected as a function of time for signal and simulated data. The parameters used in the model are $\beta = 1.4$ and $\gamma = 0.3$. The initial values of the S, I and R compartment are 9980, 2 and 0 respectively. The data shown was simulated using equation (3.5), with the underlying signal shown in the figure. Gaussian noise was introduced to the parameters and the final values of the signal.

Figure (3.3) shows a 2D marginalised posterior distribution of the two parameters γ and β produced after analysing the optimal chain. The contour shows the 68% and 95% confidence limits from the marginalised distributions. The shading shows the mean likelihood of the samples, with darker shading in the figure corresponding to sets of parameters that produce a model with a higher likelihood. The figure shows that the parameters have a clear maximum likelihood region, identified by the progressively less shaded regions around the centre.

Figure (3.4) shows the 'Triangle' plot of the parameters γ and β, produced by analysing the optimal chain. The 1D and 2D plots show the marginalised distributions of the parameters. Figure (3.4) also shows the individual marginalised distributions, highlighting that both γ and β have a shape like a Gaussian distribution, with a clear maximum. The mean and standard deviation are shown in table (3.1).

Parameter	Mean and Std
β	1.388 ± 0.033
γ	0.3 ± 0.024

Table 3.1: Table showing the SIR models marginalised mean parameter values for the simulated data.

Using the marginalised mean of each parameter, which forms the set of parameters that produces a model that best fits the data, we compare the infected cases produced by the MCMC parameters to the original parameters that defined the signal.

Figure (3.5) shows the infected compartment from the SIR model developed using the best-fit parameters found through MCMC sampling. The figure shows the original signal and the simulated data. The best-fit parameters are not identical to those used to develop the signal;

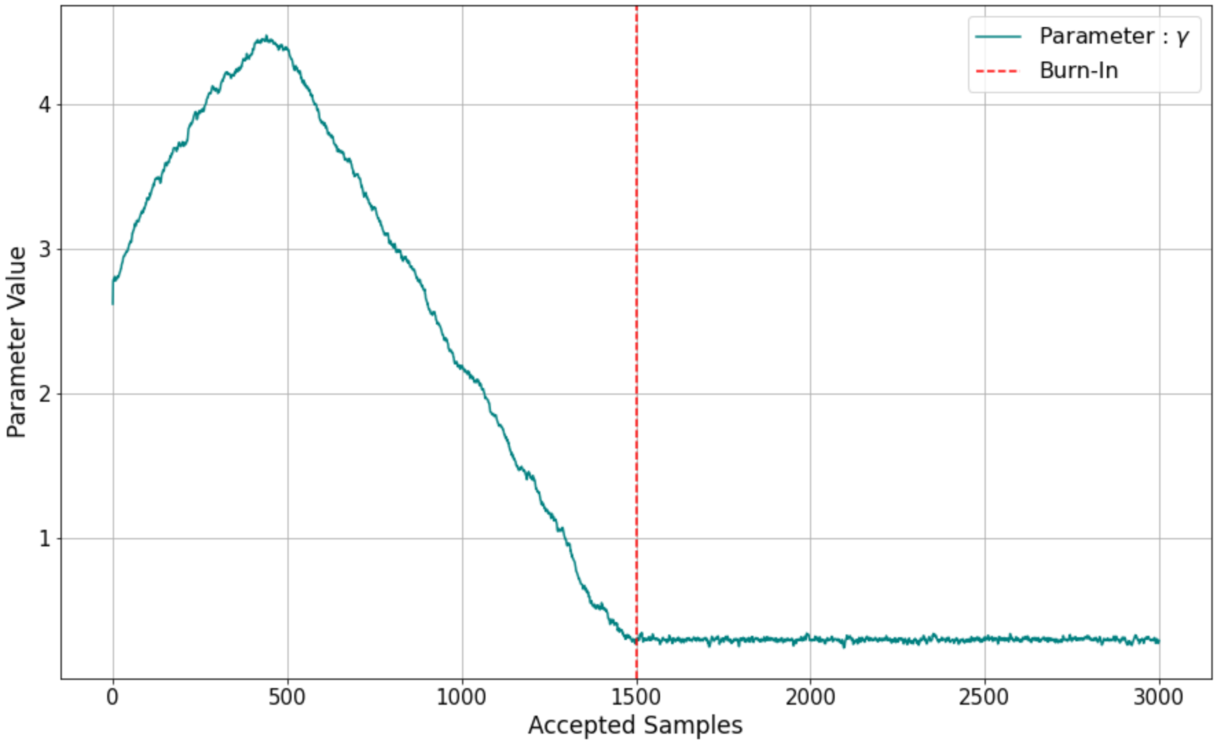

Figure 3.2: Figure showing accepted values of the parameter γ in the optimal chain found during an MCMC search. The MCMC algorithm was run for 16,000 iterations and this figure shows a subset of the accepted values. The Burn-In line highlights the point after which the MCMC algorithm begins sampling the parameter space close to the global minimum. The sampling near the global minimum is identified due to the values of γ fluctuating about a constant point.

however, they fall within a standard deviation. Consequently, the resulting Infected MCMC curve is not identical to the Infected signal curve.

Figure (3.6) shows the S, I and R compartment curves for the signal parameters $\beta = 1.4$ and γ = 0.3 and the MCMC parameters $\beta = 1.388$ and $\gamma = 0.3$. The figure shows that although the compartments S and R were not used during the MCMC search, the resulting curves produced by the best-fit parameters are almost identical to the signal curves. This indicates that the MCMC algorithm was reasonably successful in finding the parameters that best fit the simulated data.

3.3.2 Applying SIHR Compartmental Model

We define an SIHR model with the parameters $\beta = 1.4$, $\gamma = 0.3$, $\mu = 0.1$ and $\epsilon = 0.5$. Susceptible, infected, hospitalised and recovered compartments have initial population sizes of 9980, 2, 0 and 0 respectively. However, we only use the data for the infected and hospitalised compartment in the MCMC process. We are using the hospitalised compartment data in addition to the infected compartment data. This is because as models became more complicated, issues of degeneracy begin to arise. This is where, due to insufficient constraints being introduced into finding the optimal parameters, unsatisfactory results across the whole range of compartments were produced. Due to increasing the number of parameters the dimension of the parameter space increased, resulting in our developed MCMC algorithm becoming inefficient. This was made clear when sampling the 4D parameter space of the SIHR model, required upwards of 100,000 samples and multiple hours to find sets of parameters that produced satisfactory results. This therefore motivated the use of the PYMC3 package. The infected and hospitalised data and the underlying signal are shown in figure (3.7).

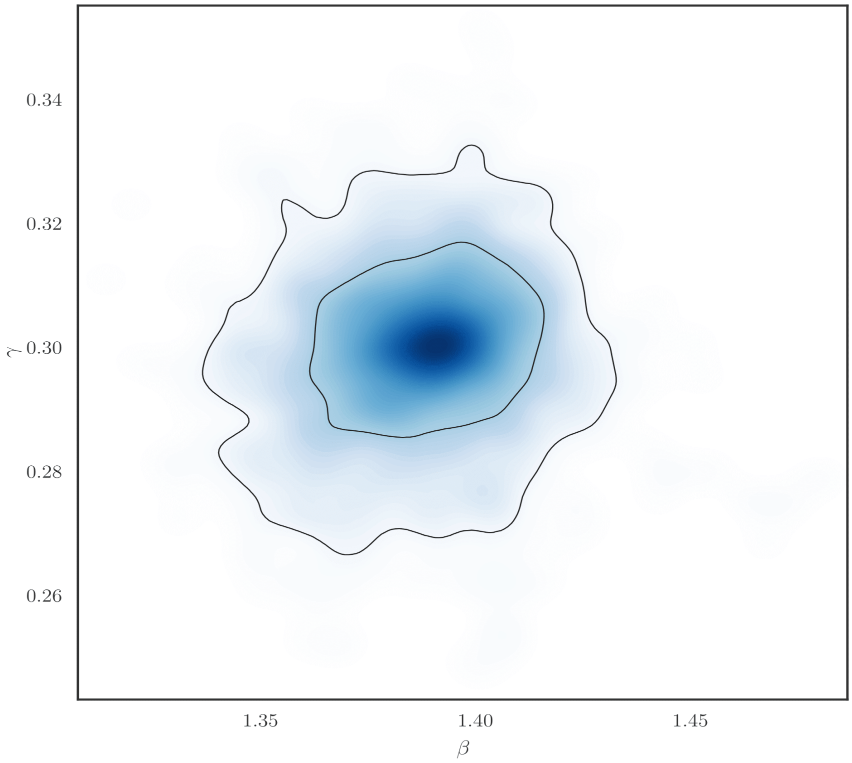

Figure 3.3: 2D marginalised posterior distribution for the parameters γ and β, produced after analysing the optimal chain. The contour shows the 68% and 95% confidence limits from the marginalised distributions. The shading shows the mean likelihood of the samples. The darker shaded regions indicate sets of parameters that produce better fits.

Gaussian noise was introduced to the original parameters and to the final values of the signal infected and hospitalised curves, with a mean at the original signal value and a standard deviation of 0.1 of the signal value. We provide the infected and hospitalised data to the PYMC3 algorithm and use the SIHR model with the same initial conditions as used to develop the signal. We run the PYMC3 sampler for 10 chains each running for 4000 samples. PYMC3 then provides the optimal chain.

Figure (3.8) shows the 'Triangle' plot of the parameters β, γ, μ and ϵ, produced by the optimal chain when using the PYMC3 sampler on the SIHR simulated data. The 1D and 2D plots show the marginalised distributions of the parameters. The 2D contour plots show that for each combination of parameters there is a clear maximum likelihood. The figure also shows that the 1D distribution of the parameters have a shape like a Gaussian distribution. The mean and standard deviation of the parameters are shown in table (3.2).

Parameter	Mean and Std
β	1.391 ± 0.123
γ	0.294 ± 0.065
μ	0.111 ± 0.036
ϵ	0.554 ± 0.191

Table 3.2: Table showing the SIHR models marginalised mean parameter values for simulated data.

Using the marginalised mean of each parameter, which is essentially the set of parameters that produces a model that best fits the data, we compare the infected and hospitalised curves produced by the parameters found during the MCMC search to the original parameters that defined the signal.

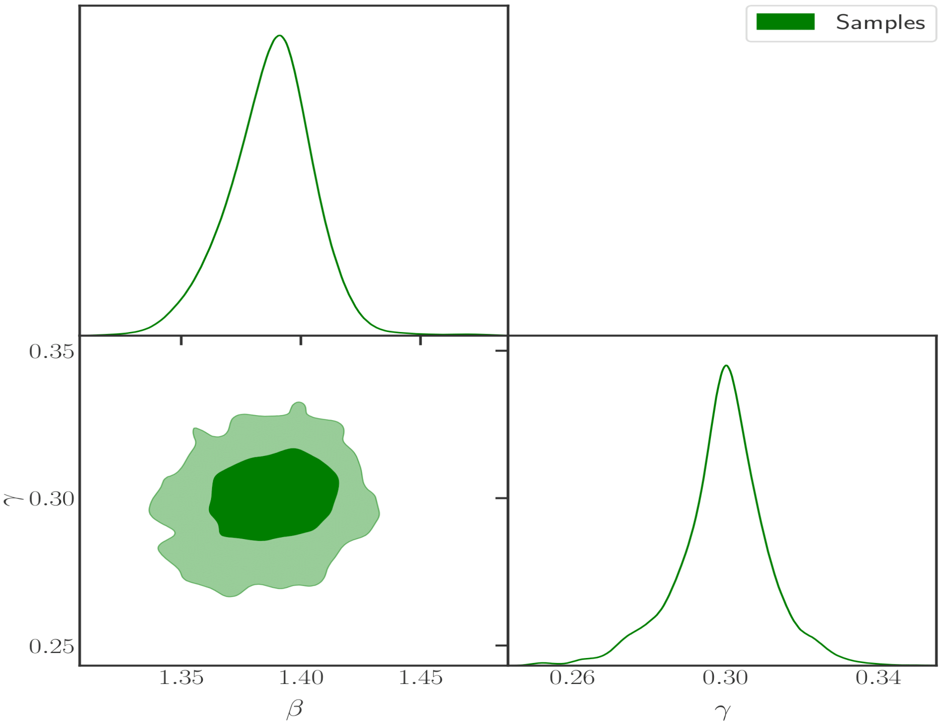

Figure 3.4: 'Triangle' plot of the parameters γ and β samples, developed by analysing the optimal chain. The 1D and 2D plots show the marginalised distributions of the parameters. The 2D contours show the 68% and 95% confidence limits. The shading shows the mean likelihood of the samples.

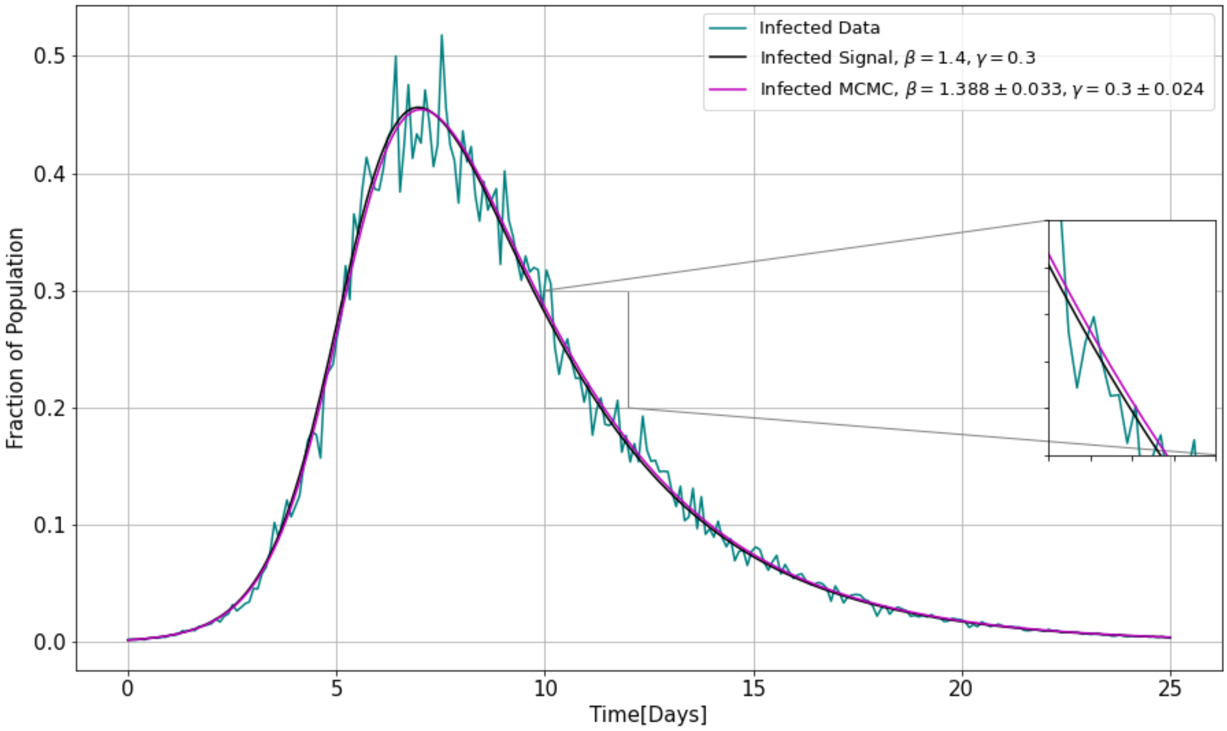

Figure 3.5: SIR compartmental model plot, showing the fraction of the population that is infected as a function of time. The figure shows curves produced by the signal, data and MCMC methods. The figure builds on figure (3.1). The curve Infected MCMC is produced by passing the SIR compartmental model the parameters $\beta = 1.388$ and $\gamma = 0.3$, which were the best fit set of parameters found during MCMC sampling.

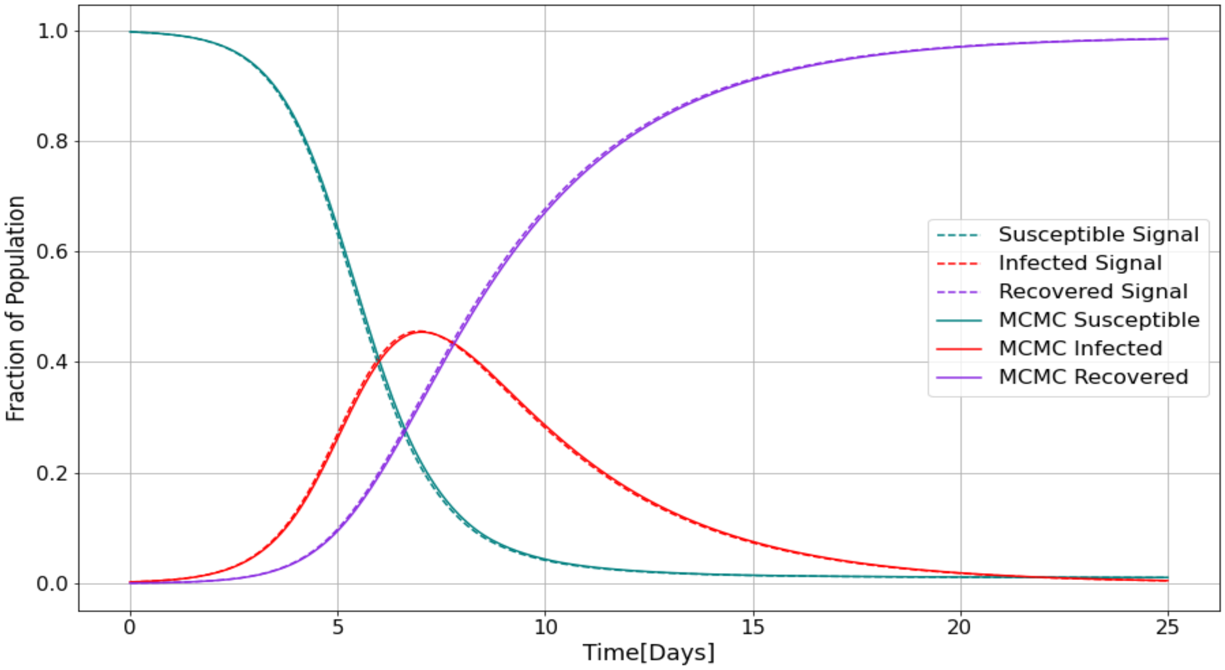

Figure 3.6: SIR compartmental model plot, showing the fraction of the population of each compartment as a function of time. The figure shows curves produced by the signal and the MCMC methods. The parameters used to develop the signal are $\beta = 1.4$ and $\gamma = 0.3$. The MCMC curves are produced by passing the SIR compartmental model the parameters $\beta = 1.388$ and $\gamma = 0.3$, which were the best fit set of parameters found during the MCMC sampling.

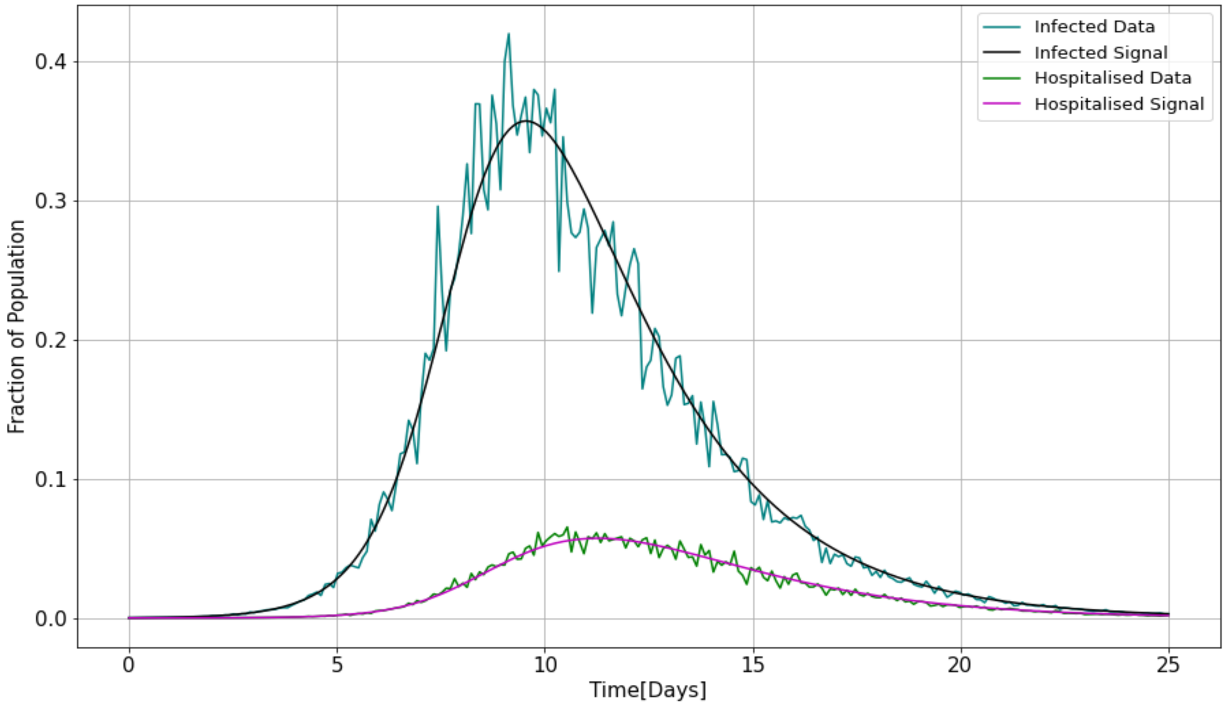

Figure 3.7: SIHR compartmental model plot, showing the fraction of the population that is infected and hospitalised as a function of time for signal and simulated data. The parameters used in the model are $\beta = 1.4$, $\gamma = 0.3$, $\mu = 0.1$ and $\epsilon = 0.5$. The initial values of the S, I, H and R compartments are 9980, 2, 0 and 0 respectively. The data was simulated using equation (3.5), with the underlying signal shown in the figure. Gaussian noise was introduced to parameters and to the final values of the signal.

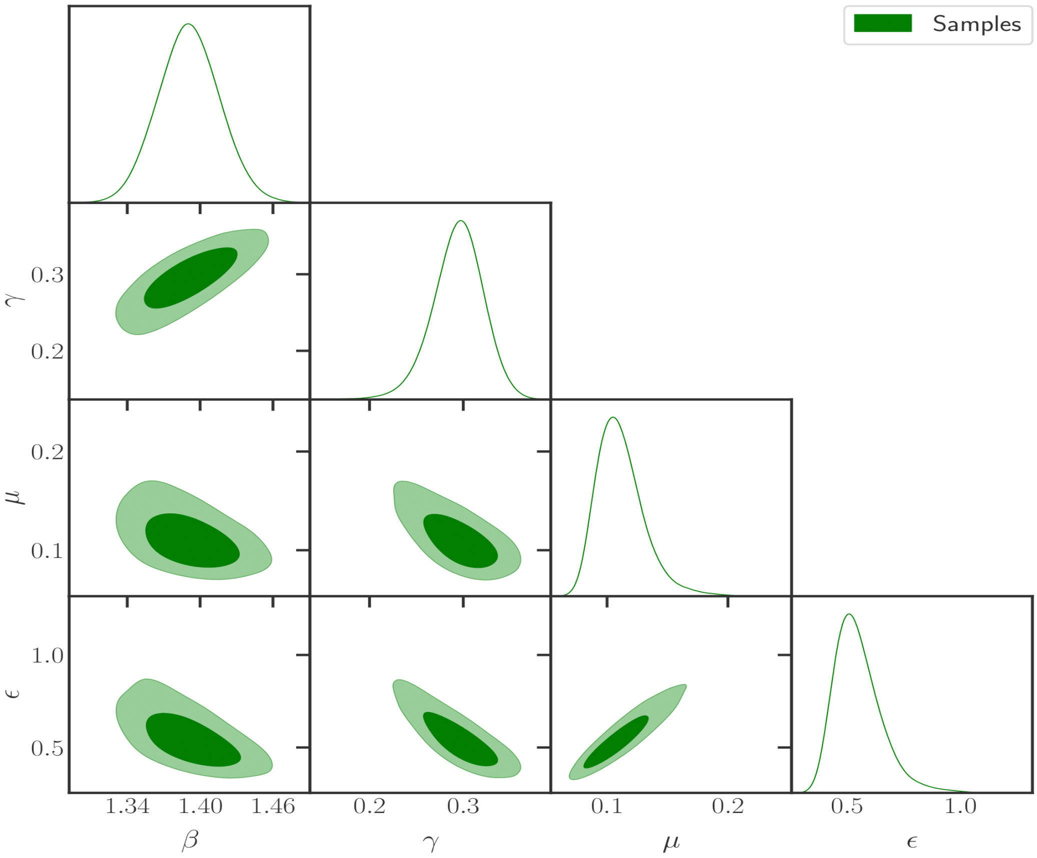

Figure 3.8: 'Triangle' plot of the parameters β, γ, μ and ϵ samples, developed by analysing the optimal chain produced by PYMC3 sampling. The 1D and 2D plots show the marginalised distributions of the corresponding parameters. The 2D contours show the 68% and 95% confidence limits. The shading shows the mean likelihood of the samples. The darker shaded regions show the set of parameters that produce better fits.

Figure (3.9) shows the infected and hospitalised compartments from the SIHR model developed using the best-fit parameters. The figure also shows the original signal and the simulated data. The best-fit parameters are not the same as those used to develop the signal, however, they fall within a standard deviation. Consequently, the resulting infected and hospitalised MCMC curves are not identical to their corresponding signals.

Figure (3.10) shows the S, I, H and R compartment curves for the signal parameters $\beta = 1.4$, $\gamma = 0.3$, $\mu = 0.1$ and $\epsilon = 0.5$ and the MCMC parameters $\beta = 1.391$, $\gamma = 0.294$, $\mu = 0.111$ and $\epsilon = 0.55$. The figure shows that although the compartments S and R were not used during the MCMC search, the best fit parameter curves are almost identical to the signal curves. Both methodologies used were able to find sets of parameters that would produce a model that best fits the simulated data.

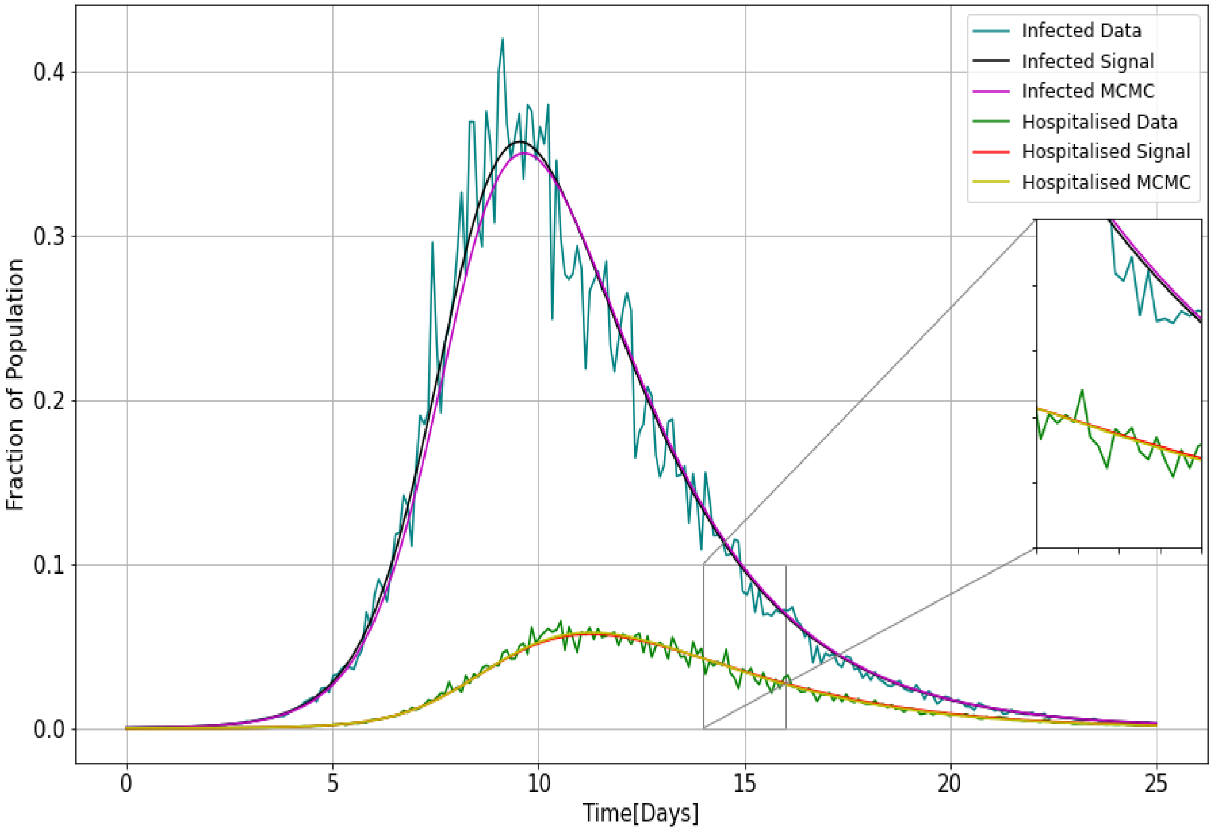

Figure 3.9: SIHR compartmental model plot, showing the fraction of the population that is infected and hospitalised as a function of time. The figure shows curves produced by the signal, data and PYMC3 sampling methods. The figure builds on the figure (3.7). The curves, Infected MCMC and Hospitalised MCMC, are produced by passing the SIHR compartmental model the parameters $\beta = 1.391$, $\gamma = 0.294$, $\mu = 0.111$ and $\epsilon = 0.55$, which were the best fit set of parameters found during the MCMC sampling.

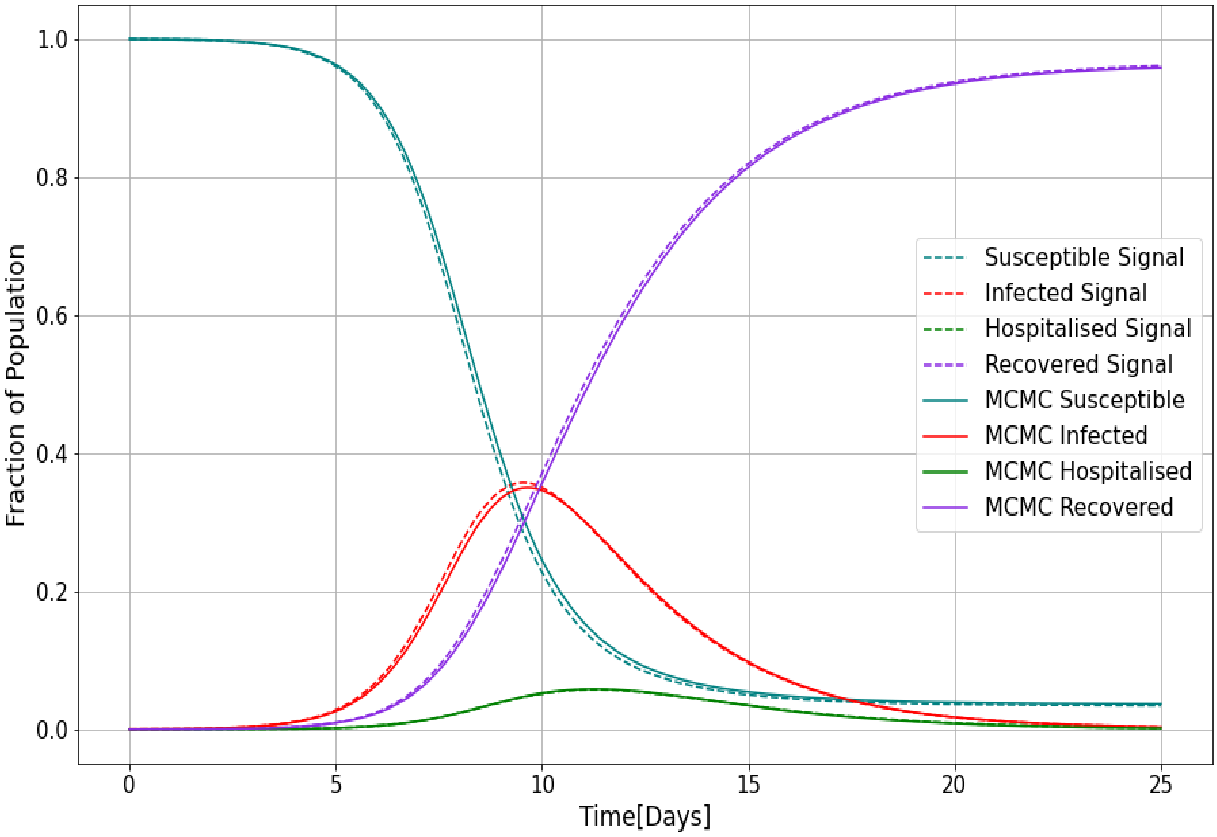

Figure 3.10: SIHR compartmental model plot, showing the fraction of the population of each compartment as a function of time. The figure shows curves produced by the signal, and the PYMC3 parameters. The parameters used to develop the signal are $\beta = 1.4$, $\gamma = 0.3$, $\mu = 0.1$ and $\epsilon = 0.5$. The MCMC curves are produced by passing the SIHR compartmental model the parameters $\beta = 1.391$, $\gamma = 0.294$, $\mu = 0.111$ and $\epsilon = 0.55$, which were the best fit set of parameters.

Chapter 4

Investigating COVID-19 Data

In this chapter. we utilise PYMC3, and the methodologies described in chapter 3 to fit the SIR and SIHR compartmental models to data on the UK COVID-19 pandemic from the period of July 2020 to January 2021.

4.1 Methodology

In this section we discuss the preparation of the COVID-19 data and how we adjusted the methodologies described in chapter 3 to fit the models and predict how the pandemic will evolve.

4.1.1 Extracting COVID-19 Data

The data used in the project was accessed from the coronavirus in the UK dashboard [29]. This is an interactive dashboard that provides an up-to-date summary of key information about the COVID-19 pandemic in the UK. The dashboard provides several key metrics for modelling purposes that can be accessed via an API. The metrics we use in our project are the daily new Infection cases and confirmed Hospital cases for England.

We generate the API on the UK government website, by requesting the metrics we want to analyse, the location (England) and the period we want to study (July 2020 - January 20201). The generated API is then accessed in Python. Each metric data contains the value of the metric at each day for the specific period.

Before we could fit the compartmental models to the data, we had to go through several steps to process the data.

1. Inspect data
    - We inspect each metric, looking for incorrect dates and missing values.
2. Clean data
    - Compartmental models are continuous and in the methodology, we compare the value in the compartment at each day to the corresponding value in the data. Therefore, having missing values would introduce a host of computational errors.
    - We replace days with no values with the value of the closest available date.
3. Rolling average

- Data has inconsistencies and noise which makes analysis difficult.
- We smooth the dataset by performing a 7 day rolling average.

We apply the methodology above to each metric.

4.1.2 Developing Markov Chain Monte Carlo

The data (metrics) we use for the SIR model is the daily Infectious cases. For the SIHR we use the daily Infectious cases and Hospitalised cases. An objective was to use the best fit parameters and resulting curves to predict how the pandemic will evolve. We define the fitting period for each of the models from 07/07/2020 to 06/11/2020. We define the predicting period from 06/11/2020 to 31/12/2020. The fitting period is the section of each of the metrics that we use as data for the MCMC algorithm. Essentially MCMC finds the parameters that best fit the data in the fitting period. We then investigate how the resulting curve behaves after the fitting period. We then compare the predictions to the actual data. We use the value of the metrics on the first day of the fitting period as the initial conditions of our compartments.

We initially investigate the SIR model with constant parameters, meaning we have a single set of parameters that define the evolution of the SIR compartments over the period. However, the pandemic is ever-evolving and so we also model the pandemic using time-varying parameters. This is where the fitting period is split into smaller periods, where each sub-period has a separate set of optimal parameters. This improves future predictions as predictions are calculated based on the current conditions of the pandemic. We adjust the methodology to incorporate time-varying parameters by initialising a set of parameters for each sub-period using a uniform distribution in PYMC3. However, we only vary and use the set of parameters when we are in their respective periods. We apply the time-varying parameters method to the SIR and SIHR models.

4.2 Results and Discussion

In this section, we discuss the results of applying the SIR and SIHR models to the COVID-19 pandemic data for the period 07/07/2020 to 31/12/2020.

4.2.1 COVID-19 Data

The data used in this project is shown in figure (4.1).

The metrics: daily new Infection cases and Hospitalised cases are shown in figure (4.1). The Cut-Off line indicates the period after which we aim to predict the behaviour of the metrics.

The metric daily new Infection cases is not the same as the infected compartment, as the latter represents the total number of infected individuals. To make the compartmental model compatible with the data, we convert the infected compartment to daily number of new infected individuals using the equation below,

$$I_{Daily} = \frac{\beta S I}{N} \tag{4.1}$$

where I_{Daily} is the daily new infection cases. Note this is the same as the RHS of $\frac{dS}{dt}$ in equation (2.1).

Although in this project we aim to describe the evolution of the pandemic based only on the compartmental models, it will be useful to understand why we decided to predict this period of

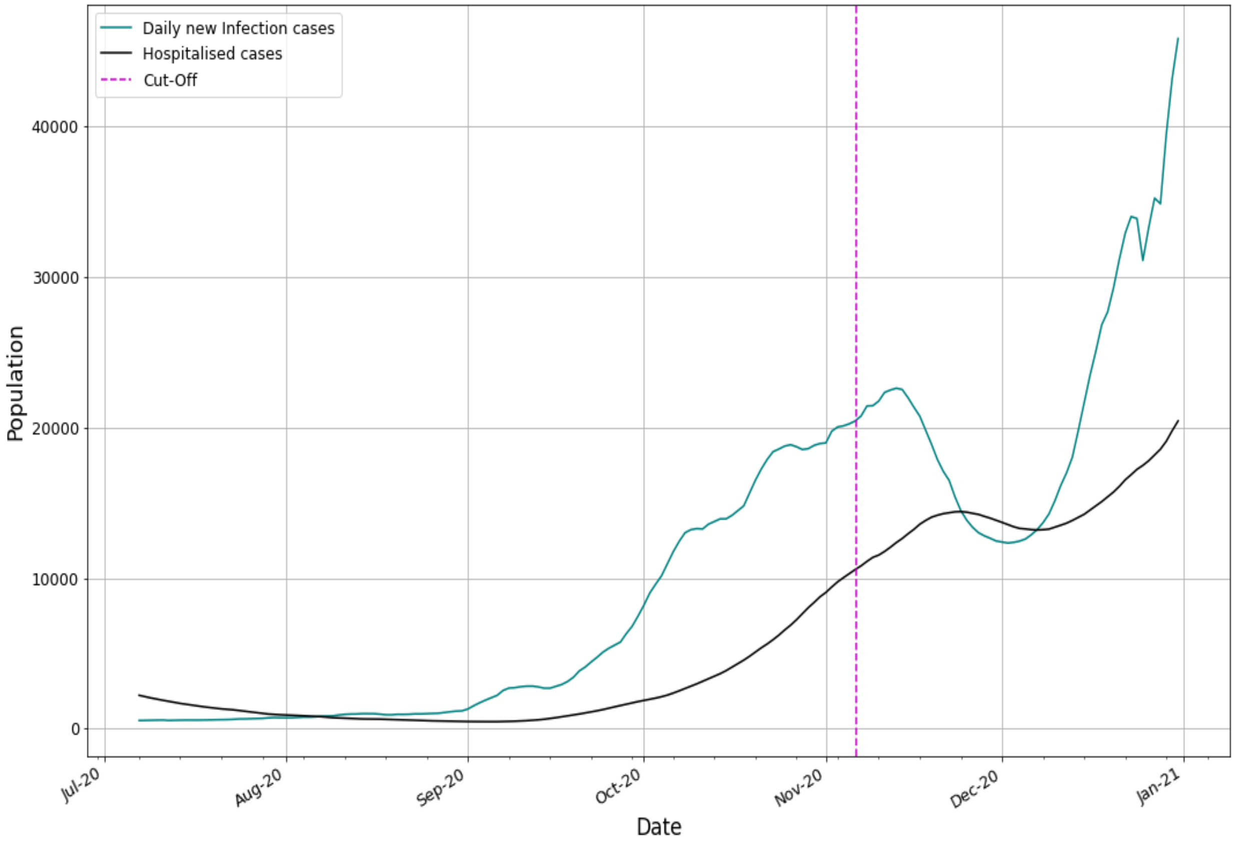

Figure 4.1: Figure showing England's COVID-19 pandemic data for the period 07/07/2020 to 31/12/2020. Figure shows the metrics daily new Infection cases and Hospitalised cases as a function of time. The Cut-Off line indicates the period after which we aim to predict the behaviour of the individual metrics.

the pandemic. A key reason is that prior to this period, the way the data was being collected, processed, and provided to analysts was haphazard and had a host of inconsistencies, making modelling the pandemic challenging. In addition to this, the data was not well understood as the pandemic was relatively recent. We also selected this period as it was the beginning of a new mini pandemic. The restriction levels (isolation and contact) were relatively fixed during this period and the new Alpha variant of the virus began to overtake the original wild variant. The combination of these factors meant that this was an ideal period to apply the methods we have discussed.

4.2.2 Applying SIR Compartmental Model

We initially model the fitting period using a single set of parameters. We then specify initial conditions, using the number of daily new infected individuals on the date 07/07/2020, $I_{Daily} = 552$. We then assume S, the initial population of England as N_{Eng}, minus I_{Daily}, which is $S = 53,999,448$. We also assume that there are no recovered individuals. We produce 10 chains each running for 4000 iterations. We only use the metric daily new Infection cases as data, as we are using the SIR model. The distribution of parameters is shown in figure (4.2).

Figure (4.2) shows the 'Triangle' plot of the parameters γ and β produced by the optimal chain when performing MCMC sampling on the fitting period for the daily Infection cases data. The figure shows that the individual marginalised distributions of the parameters follow a Gaussian distribution, the mean and standard deviation of the parameters are shown in table (4.1).

The reproductive number is $R_0 = 1.01$. Using the marginalised mean of each parameter, we

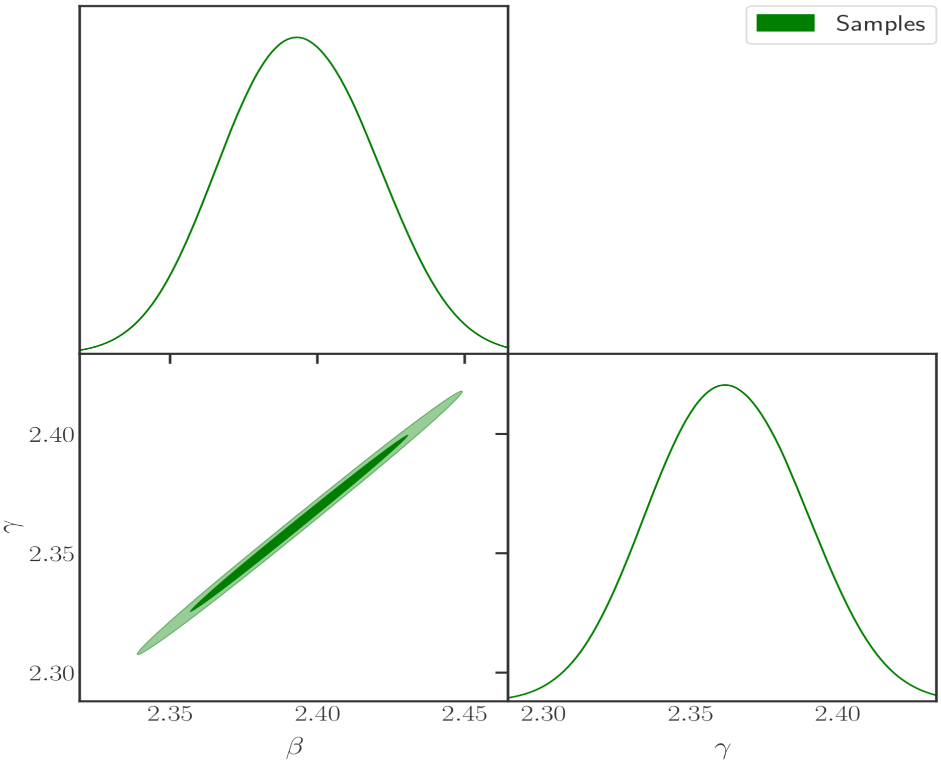

Figure 4.2: 'Triangle' plot of the parameters γ and β samples, determined by performing MCMC sampling on the daily Infection cases data in England for the COVID-19 pandemic, for the period 07/07/2020 to 06/11/2020 . The 1D and 2D plots show the marginalised distributions of the corresponding parameters. The 2D contours show the 68% and 95% confidence limits. The shading shows the mean likelihood of the samples.

Parameter	Mean and Std
β	2.394 ± 0.043
γ	2.362 ± 0.043

Table 4.1: Table showing the SIR models marginalised mean parameter values for the real data fitting period, 07/07/2020 to 06/11/2020.

can plot our develop daily new Infection cases model and compare it to the actual data, shown in figure (4.3).

Figure (4.3) shows the fitted SIR model for the best fit parameters $\beta = 2.394$ and $\gamma = 2.362$. The figure only shows the infected compartment, having been adjusted to show the daily new Infection cases. The predicted daily new Infection cases shows the predicted case numbers. The results show that during the fitting period the curve is not able to fit the data properly, this is to be expected as we are only considering one set of parameters. As a result, during the fitting period, we overestimate the case numbers for September and underestimate in November. The prediction made by the curve is significantly lower than the actual case numbers. However, one thing the prediction does get correct is that at least for two weeks after the prediction period, the case numbers will decrease. After that, however, the epidemic evolves, and the model is not able to predict this.

Constant parameter models have poor performance in modelling the pandemic. Therefore, we explore the use of time varying parameters. We split the fitting period (before Cut-Off) into 5 equal periods. Using the SIR model, we obtain a set of parameters for each period and use the final set of parameters (the period just before Cut-Off) as the parameters to predict the metrics. Using the same initial conditions and the daily new Infection cases data, we sample

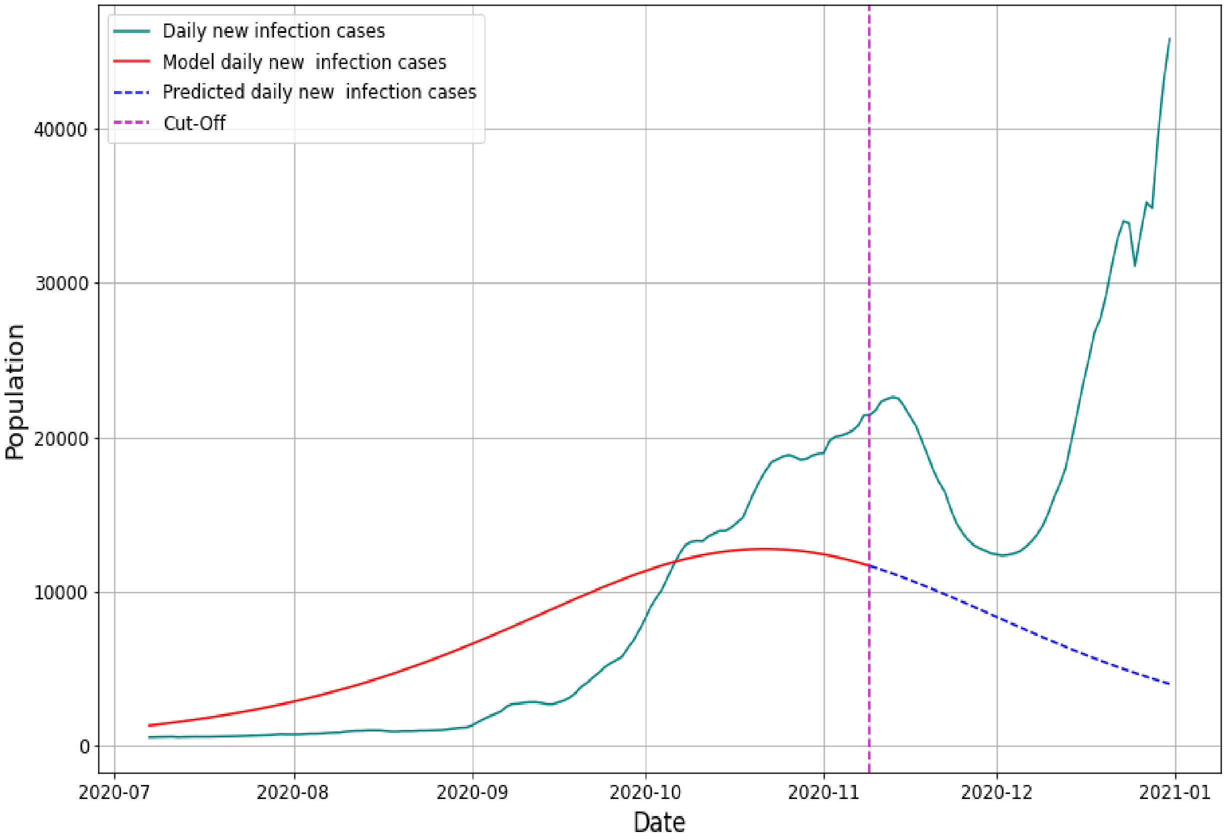

Figure 4.3: Fitting of the SIR compartmental model to the metric daily new Infection cases. The Cut-Off line indicates the period after which we predict the behaviour of the metric. The Model and predicted daily new Infection curves were generated using the best fit parameters $\beta = 2.394$ and $\gamma = 2.362$.

with 4 chains each running for 1000 iterations. The distribution of the final set of parameters are shown in figure (4.4).

Figure (4.4) shows that the individual marginalised distributions for the parameters follow a Gaussian distribution, however the standard distribution are significantly smaller than the other parameters seen so far. Also, unlike the circular 2D distributions seen in chapter 3, here they are elongated, showing that there is a series of optimal parameter sets. The mean and standard deviation of the parameters for each period during the fitting phase is shown in table (4.2).

Periods	β	γ
Period 1	2.5023 ± 0.0002	2.5001 ± 0.0002
Period 2	2.5034 ± 0.0002	2.5014 ± 0.0003
Period 3	2.5261 ± 0.0003	2.4982 ± 0.0002
Period 4	2.5223 ± 0.0002	2.4419 ± 0.0003
Period 5	2.5348 ± 0.0002	2.5042 ± 0.0001

Table 4.2: Table showing the SIR models marginalised mean parameter values for the individual periods during the real data fitting period, 07/07/2020 to 06/11/2020.

Using the marginalised mean of each parameter, we plot the fitted model and compare it to the data on the daily new Infection cases, shown in figure (4.5).

The results in figure (4.5) show that introducing time varying parameters significantly improves the ability for compartmental models to fit the data. Although we define 5 periods, this can be

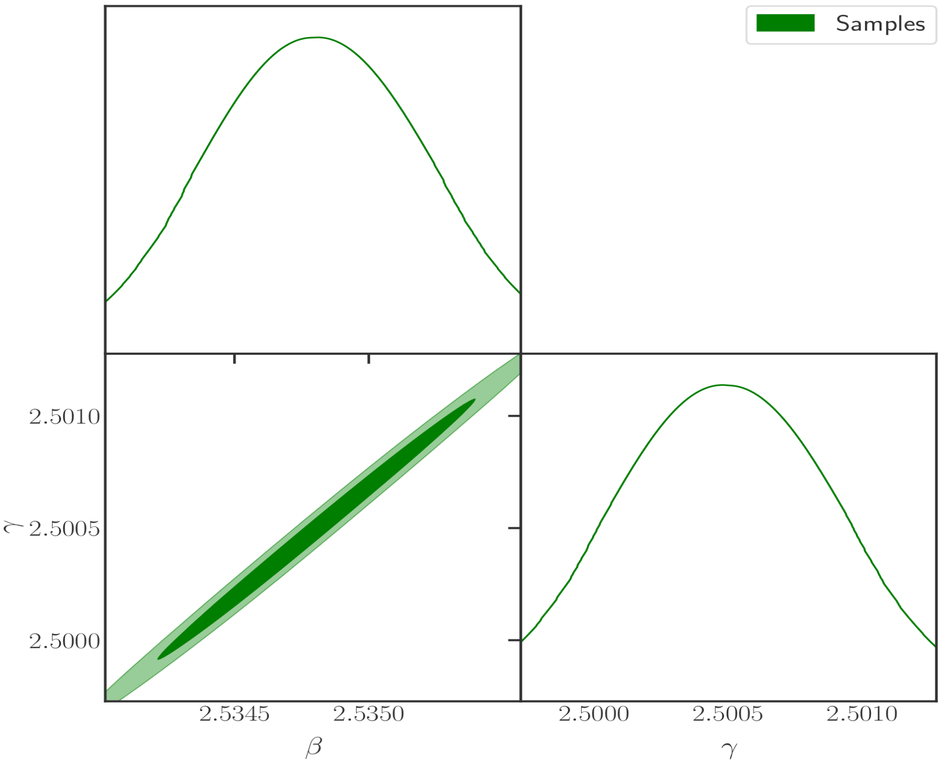

Figure 4.4: ‘Triangle’ plot of the parameters γ and β samples for the final period of the fitting period of the data, determined by performing MCMC sampling on the daily new Infection cases data in England for the COVID-19 pandemic, for the period 07/07/2020 to 06/11/2020 . The 1D and 2D plots show the marginalised distributions of the corresponding parameters. The 2D contours show the 68% and 95% confidence limits. The shading shows the mean likelihood of the samples.

increased to obtain better fits. However, as we increase the number of periods, we may end up over fitting the data. The prediction performs significantly better than the constant parameter method shown in figure (4.3), with the predictions more closely following the actual data for up to three weeks. Although as we approach late December the model is not able to predict the spike in cases. However, if the method described here is constantly employed on a daily basis, essentially moving the fitting period up a everyday then as cases begin to rise the model will be able to account for this change in infection cases.

4.2.3 Applying SIHR Compartmental Model

In this section we discuss the results of applying the SIHR model to the COVID-19 pandemic data. We split the fitting period into three periods and obtain a set of parameters for each period. We set the initial $I_{Daily} = 552$ and the initial number of hospitalised individuals to, $H = 2224$. We assume initial S is, $S = N_{Eng} - I_{Daily} - H$. We assume there are no recovered individuals. We run the MCMC sampler with 8 chains each running for 2000 iterations.

Figure (4.6) shows the distribution of the parameters for the final fitting period of the SIHR model. The figure shows that the individual marginalised distributions for the parameters have a shape that is like a Gaussian distribution, with a clear maximum. Also, unlike the circular 2D distributions seen in chapter 3, here parameters μ and ϵ have an elongated distribution, indicating strong correlation between the two parameters. The same can also be said about β and γ.

The mean and standard deviation of the parameters for each period during the fitting period is shown in table (4.3).

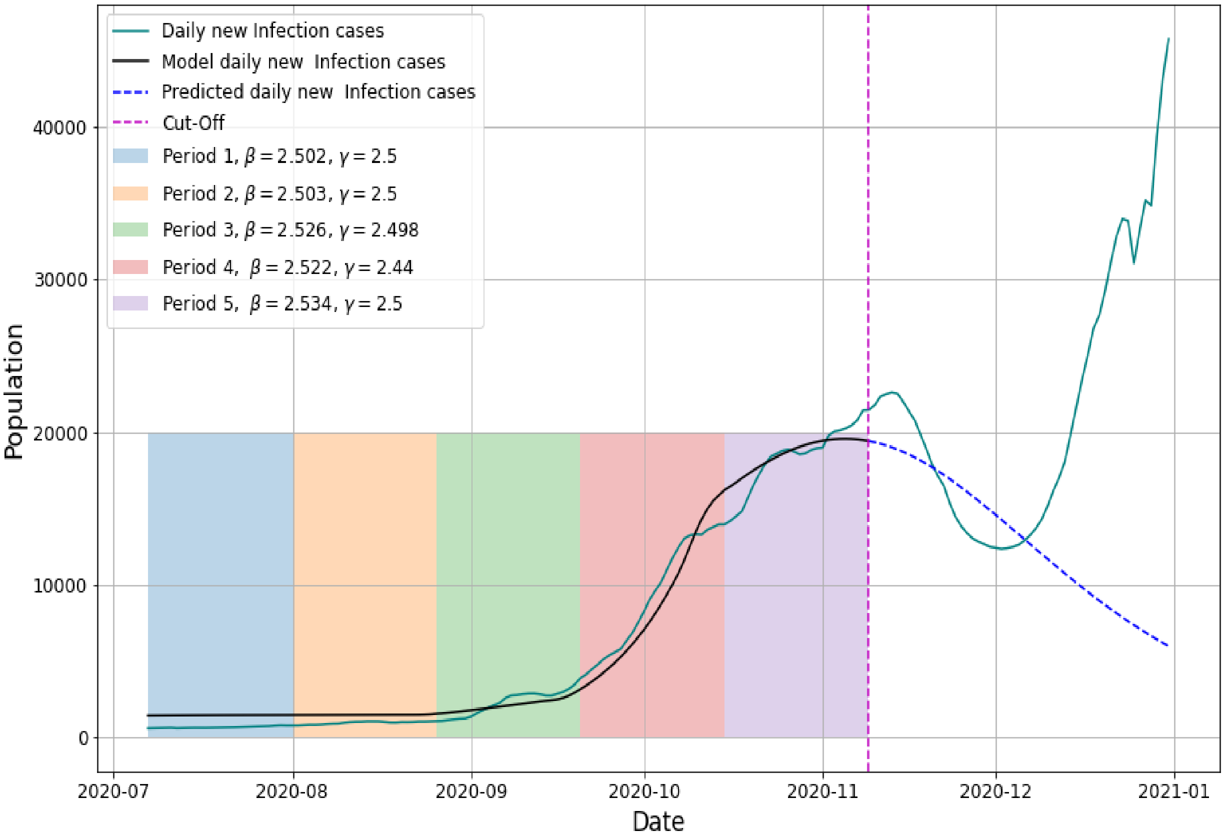

Figure 4.5: Fitting of the SIR compartmental model for the metric daily new Infection cases. The Cut-Off line indicates the period after which we predict the behaviour of the metrics. Each highlighted box shows an individual fitting period and the corresponding parameters found during the MCMC search. The predicted daily new Infection curve was generated using the best the fit parameters found for Period 5.

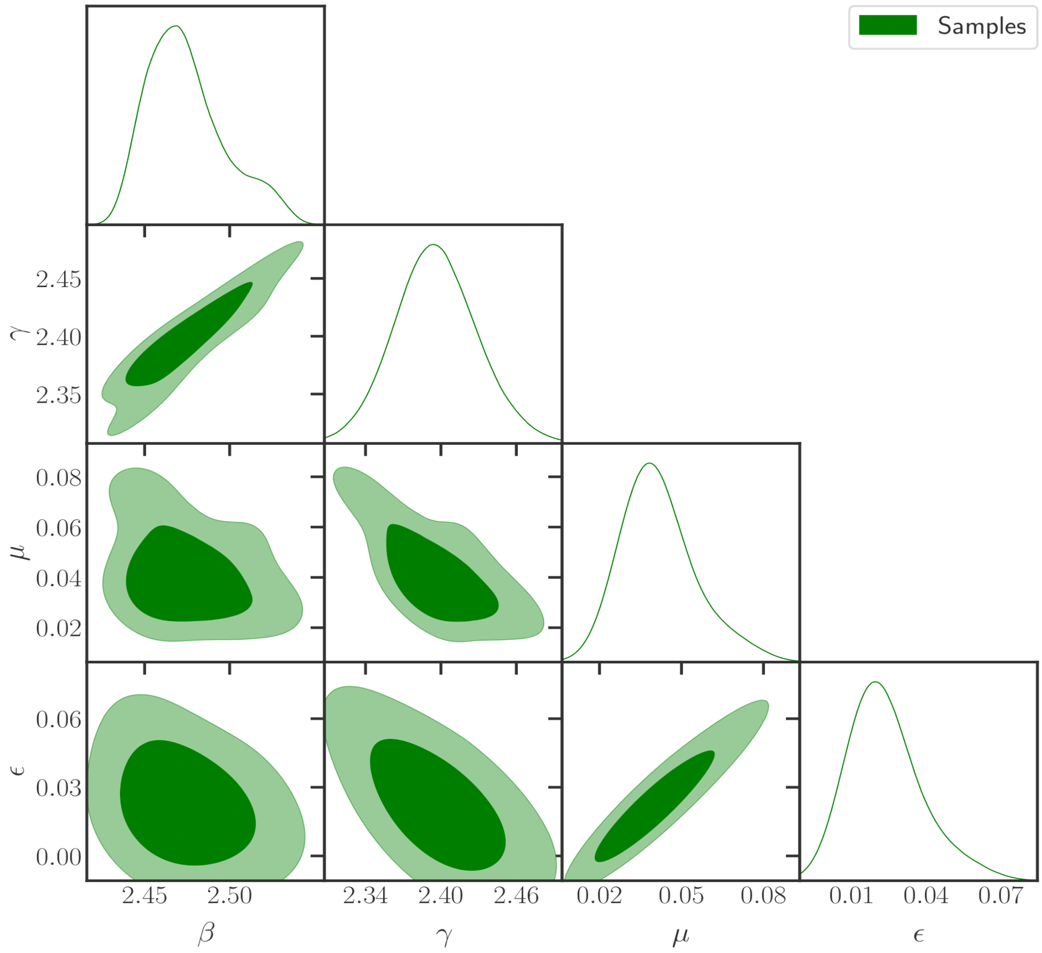

Figure 4.6: 'Triangle' plot of the parameters β , γ, μ and ϵ for the final fitting period of the data, for the SIHR compartmental model. Distributions were determined by performing MCMC sampling on the daily infection cases data and hospitalised cases in England for the COVID-19 pandemic, for the period 07/07/2020 to 06/11/2020. The 1D and 2D plots show the marginalised distributions of the corresponding parameters. The 2D contours show the 68% and 95% confidence limits. The shading shows the mean likelihood of the samples.

Parameters	Period 1	Period 2	Period 3
β	2.219 ± 0.062	2.647 ± 0.048	2.474 ± 0.051
γ	2.183 ± 0.055	1.965 ± 0.063	2.396 ± 0.068
μ	0.033 ± 0.025	0.588 ± 0.052	0.041 ± 0.031
ϵ	0.051 ± 0.035	1.051 ± 0.034	0.023 ± 0.033

Table 4.3: Table showing the SIHR models marginalised mean parameter values for the individual periods during the real data fitting period, 07/07/2020 to 06/11/2020.

Using the marginalised mean of each parameter, we plot the fitted model and compare it to the data on the daily new Infection cases and the Hospitalised cases, shown in figure (4.7).

The results in figure (4.7) shows that introducing more periods improves the ability to fit models to the data. It also improves the model's capability to predict how each of the metrics will evolve. Although we only use 3 time periods, the curve Predicted daily new Infection cases performs almost as well as the one shown in figure (4.5). This could be due to splitting the fitting period based on the sharp changes in the data. Also introducing the hospitalised compartment means that the model is able model more complex dynamics. The predicted curves for the daily new Infection cases and Hospitalised cases perform well until December, where we again encounter the issue of the sharp rise in infection cases and hospitalised cases. As discussed in the SIR section, we can predict the behaviour of this later data by updating our model with a closer fitting period. A core reason it no longer predicts it well, is because the assumptions have changed. Although in this project we have not assumed any external constraint on the parameters, instead we find the optimal sets of the parameters, it may be the case that that the optimal set is not the same as the actual parameters of that period of the pandemic. A clear example is that in reality $\frac{1}{\gamma} \approx 14$ days, however for the SIHR model here in period 3, $\frac{1}{\gamma} \approx 0.417$ days. As a result, although different optimal parameter sets may produce similar fits, the reported values may be different from the actual values. This shows the limitations of just looking for the optimal set of parameters without assuming any priors and is something that can be expanded on in further work.

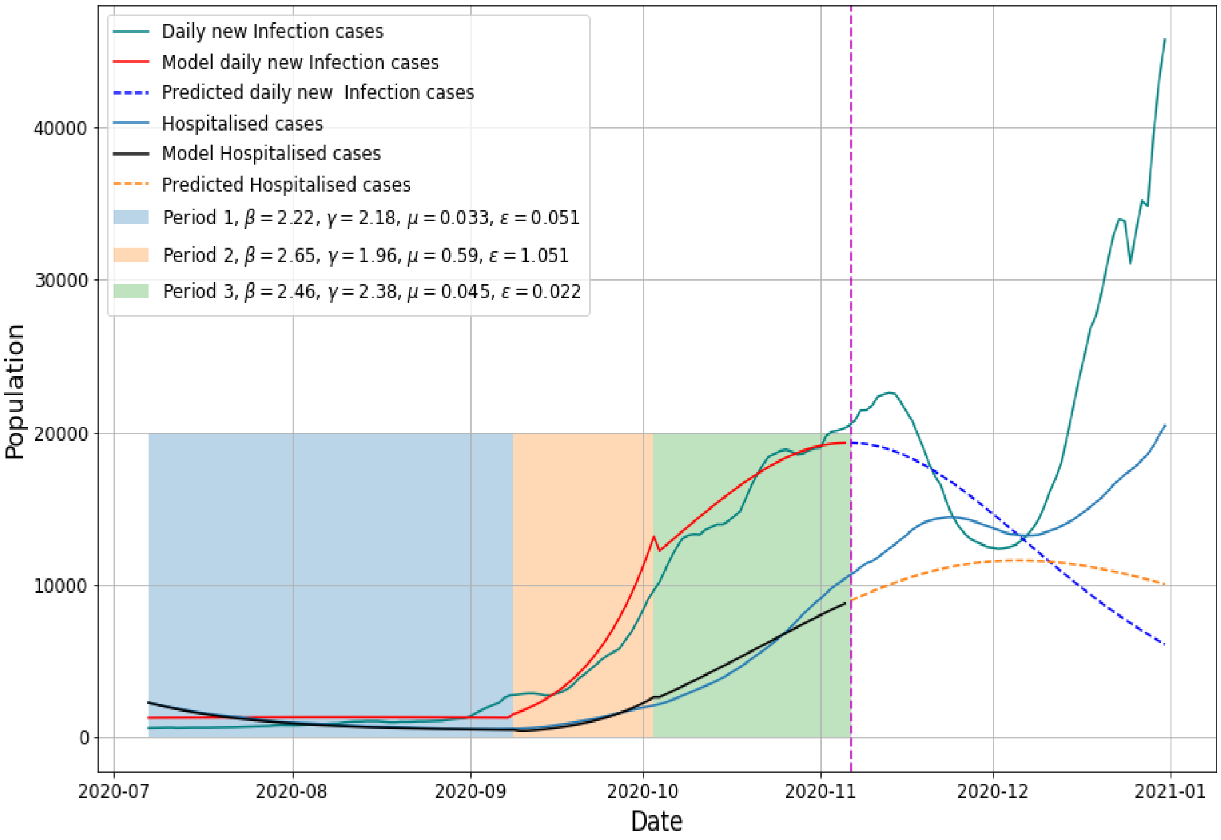

Figure 4.7: Fitting of the SIHR compartmental model to the metrics daily new Infection cases and Hospitalised cases. The Cut-Off line indicates the period after which we predict the behaviour of the metrics. Each highlighted box shows an individual fitting period and the corresponding parameters found during the MCMC search. The predicted daily new Infection cases and Hospitalised cases were generated using the best fit parameters found for Period 3.

Chapter 5

Conclusion

In this project we successfully investigate compartmental models of varying complexity by developing the SIR, SIHR and SVEASyHRD models and obtain numerical solutions to their respective ODE. We develop simulated data using the SIR and SIHR compartmental models and use our own MCMC algorithm and PYMC3 to find the parameters of the simulated data using the SIR and the SIHR model. We use the chains produced during MCMC sampling to obtain the posterior distributions of the parameters and show that MCMC sampling can successfully find the parameters of the simulated data. Finally, we use the SIR and SIHR compartmental models with PYMC3 to find the parameters of the COVID-19 pandemic for the period July 2020 to January 2021. We fit the SIR model to the data, daily new Infection cases, for the fitting period, 07/07/2020 to 06/11/2020, and obtain the parameters, $\beta = 2.394 \pm 0.043$and $\gamma = 2.362 \pm 0.043$. The resulting curves show constant parameters are not able to fit and predict the pandemic data effectively. We then use time-varying parameters for the SIR model, splitting the fitting period into 5 individual periods, determining an optimal set of parameters for each. We use the final fitting period parameters, $\beta = 2.5348 \pm 0.00023$ and $\gamma = 2.50422 \pm 0.0001$, to predict the evolution of the infection cases for the predicting period 06/11/2020 to 31/12/2020. We then use time-varying parameters with the SIHR model for the infection cases and the hospitalised cases data and split the fitting period into 3 sub-periods. We determine the final fitting period parameters, $\beta = 2.474 \pm 0.051$, $\gamma = 2.396 \pm 0.068$, $\mu = 0.041 \pm 0.031$ and $\epsilon = 0.023 \pm 0.033$ and use these to predict how the compartments will evolve. We see that for both models, time-varying parameters perform significantly better, and can predict the general trend from 06/11/2020 to 07/12/2020. However, after the date, the pandemic evolves, the underlying assumptions behind the fitted models change and we are no longer able to predict the trend correctly.

This project was an initial exploration into using compartmental models to understand the COVID-19 pandemic and we were successful in achieving the aims we set out. The next steps in expanding the work detailed in this project include investigating more complex compartmental models by for example introducing age related compartments. Further work can also be done by fitting more complex models to the pandemic data and utilising additional metrics such as confirmed deaths due to COVID-19 and evaluate the relative performance between different models. In this project we modelled the pandemic by assuming no prior information, and only use the compartmental models and the data to predict case numbers. This work can be further expanded by researching prior information on parameters such as, γ, and constraining them during MCMC sampling. This would allow us to interpret our results and provide evidence-based understanding to changes during the pandemic and real-world interpretability.

Acknowledgements

I would like to thank my supervisor, Prof. Carlo R. Contaldi for the continuous help and guidance throughout the whole project. The advice I received was invaluable, and it has been a pleasure to work with him. I also thank my project partner for the stimulating discussions and insights that were integral to this project. Finally, I would like to thank my family and friends who have supported and helped me with their valuable suggestions and feedback.

Bibliography

[1] Anthony S. Fauci and H. Clifford Lane and Robert R. Redfield. “Covid-19 — Navigating the Uncharted”. The New England journal of medicine, vol. 382, pp. 1268-1269, March 2020. https://doi.org/10.1056/NEJMe2002387.

[2] Thirumalaisamy P. Velavan and Christian G. Meyer. “The COVID-19 epidemic”. Tropical Medicine & International Health, vol. 25, pp. 278-280, March 2020. https://doi.org/10.1111/tmi.13383.

[3] Li, Qun et al. “Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia”. The New England journal of medicine, vol. 382, pp 1199-1207, March 2020. https://doi.org/10.1056/NEJMoa2001316.

[4] Willem Thorbecke. “The Impact of the COVID-19 Pandemic on the U.S. Economy: Evidence from the Stock Market”. Journal of Risk and Financial Management, vol. 13, pp.233, October 2020. https://doi.org/10.3390/jrfm13100233.

[5] Jackson, James K. et al. “Global Economic Effects of COVID-19”. https://crsreports.congress.gov - Accessed 03/04/2022

[6] “COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)”. Johns Hopkins University. - Accessed 20/08/2021

[7] Nadia Akseer and Goutham Kandru and Emily C Keats and Zulfiqar A Bhutta. “COVID-19 pandemic and mitigation strategies: implications for maternal and child health and nutrition”. The American journal of clinical nutrition, vol. 112 , pp. 251-256, August 2020. https://doi.org/10.1093/ajcn/nqaa171.

[8] D. J. Daley and J. Gani. “Epidemic Modelling”. Cambridge University Press, February 1984. https://doi.org/10.1017/CBO9780511608834.

[9] Gilberto M. Nakamura et al. “Efficient method for comprehensive computation of agent-level epidemic dissemination in networks”. Scientific reports, vol. 7, pp.40885, February 2017. https://doi.org/10.1038/srep40885.

[10] Adam D. “Special report: The simulations driving the world’s response to COVID-19”. Nature, vol. 580, pp. 316–318. https://doi.org/10.1038/d41586-020-01003-6.

[11] W O Kermack and A G Mckendrick.“A Contribution to the Mathematical Theory of Epidemics”. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, vol. 115, pp. 700-721, 1927.

[12] Roberto N Padua and Alfeo B Tulang. “A Density–Dependent Epidemiological Model for the Spread of Infectious Diseases”. Liceo Journal of Higher Education Research, vol.6, December 2010. https://doi.org/10.7828/ljher.v6i2.62.

[13] Yang, W. et al. “Rational evaluation of various epidemic models based on the COVID-19 data of China”. March 2020. http://arxiv.org/abs/2003.05666.

[14] Hethcote H. “The Mathematics of Infectious Diseases”. SIAM Review, vol. 42, pp. 599-653, January 2000. https://doi.org/10.1137/S0036144500371907.

[15] Becker NG et al. “The reproduction number using Mathematical Models to Assess Responses to an Outbreak of an Emerged Viral Respiratory Disease”. National Centre for Epidemiology and Population Health, 2006. ISBN 1-74186-357-0.

[16] P. Fine,, K. Eames and D. L. Heymann. “Herd Immunity: A Rough Guide”. Clinical Infectious Diseases, vol. 52, pp. 911-916, April 2011. https://doi.org/10.1093/cid/cir007.

[17] Virtanen, Pauli et al.“SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. Nature Methods, vol.17, 2020. https://doi.org/10.1038/s41592-019-0686-2.

[18] Berg, Bernd A. “Markov Chain Monte Carlo Simulations and Their Statistical Analysis”. World Scientific, 2004.

[19] Geyer, Charles J. “Practical Markov Chain Monte Carlo.” Statistical Science, vol. 7,1992. http://www.jstor.org/stable/2246094.

[20] Brooks, S. “Markov chain Monte Carlo method and its application”. Journal of the Royal Statistical Society: Series D (The Statistician), vol. 45, pp. 60-100, March 1998. https://doi.org/10.1111/1467-9884.00117.

[21] Bernardo, José M and Adrian FM Smith. “Bayesian theory”. John Wiley Sons, vol. 405, pp 109, 2009.

[22] Berger, James O and Bernardo, José M. “On the Development of the Reference Prior Method, in Bayesian Statistics”. Oxford University Press, 1992.

[23] Bridle, S. (2002). “Cosmological parameters from CMB and other data: A Monte Carlo approach”. Physical Review D - Particles, Fields, Gravitation and Cosmology, vol. 66, November 2002. https://doi.org/10.1103/PhysRevD.66.103511.

[24] Carson Chow. “Bayesian paramter estimation”. www.sciencehouse.wordpress.com/bayesian-parameter-estimation. Accessed - 10/04/2022

[25] Carson Chow. “MCMC and fitting models to data”. www.sciencehouse.wordpress.com/ mcmc-and-fitting-models-to-data. Accessed - 10/04/2022

[26] Tudor Barbu. “Variational Image Denoising Approach with Diffusion Porous Media Flow”. Abstract and Applied Analysis, 2013. https://doi.org/10.1155/2013/856876.

[27] Ott, Miles Q et al. “Bayes Rules! An Introduction to Applied Bayesian Modelling”. United Kingdom, CRC Press, 2022.

[28] Lewis, A. “GetDist: a Python package for analysing Monte Carlo samples”.2019. https://doi.org/10.48550/arXiv.1910.13970.

[29] Salvatier J., Wiecki T.V., Fonnesbeck C. “Probabilistic programming in Python using PyMC3”. PeerJ Computer Science, (2016). https://doi.org/10.7717/peerj-cs.55.

[30] UK Health Security Agency. “GOV.UK Coronavirus (Covid-19) in the UK”. https://coronavirus.data.gov.uk/. - Accessed - 5/03/2022