A Bayesian approach to the analysis of infectious disease data using continuous-time stochastic models

Petros Barmpounakis

Supervisor: Nikolaos Demiris

Department of Statistics

Athens University of Economics and Business

«The implementation of the doctoral thesis was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research" Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities».

This dissertation is submitted for the degree of

Doctor of Philosophy in Statistics

School of Information Sciences & Technology September 2023

Μια Μπεϋζιανή προσέγγιση για την ανάλυση δεδομένων μολυσματικών νοσημάτων χρησιμοποιώντας στοχαστικά μοντέλα συνεχούς χρόνου

Πέτρος Γεώργιος Μπαρμπουνάκης

Επιβλέπων: Δρ. Νικόλαος Δεμίρης

Τμήμα Στατιστικής

Οικονομικό Πανεπιστήμιο Αθηνών

«Η υλοποίηση της διδακτορικής διατριβής συγχρηματοδοτήθηκε από την Ελλάδα και την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος «Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Δια Βίου Μάθηση», *2014-2020*, στο πλαίσιο της Πράξης «Ενίσχυση του ανθρώπινου δυναμικού μέσω της υλοποίησης διδακτορικής έρευνας Υποδράση *2:* Πρόγραμμα χορήγησης υποτροφιών ΙΚΥ σε υποψηφίους διδάκτορες των ΑΕΙ της Ελλάδας».

Διατριβή για την απόκτηση

Διδακτορικού διπλώματος στη Στατιστική

Σχολή Επιστημών &

Τεχνολογίας της Πληροφορίας

Σεπτέμβριος 2023

I would like to dedicate this thesis to the memory of my beloved grandfather,

Alexios

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

September 2023

Acknowledgements

The work in this thesis was completed under the supervision of Dr. Nikolaos Demiris. I would like to wholeheartedly thank him for his guidance, support, and friendship and for showing me the way through tough times. I would not have been able to carry out this work without him and for that, I would always be grateful.

I would also like to thank Dr. Vana Sypsa, Dr. Ioannis Kontoyiannis and Dr. George Pavlakis for their contributions to Chapter 2 of this thesis. It was an insightful collaboration and I hope they enjoyed it as much as I did.

Furthermore, I am grateful to Dr. Kostas Kalogeropoulos and Dr. Petros Dellaportas for their intelligent and useful comments on Chapter 3.

At last, I would like to say a big thank you to my family, my close friends and my precious companion, my dog Rallou, for their support and for bearing with me through countless hours of moaning about the troubles that arise when finishing a PhD thesis.

I am grateful to the National Scholarship Foundation (IKY) for founding a part of this work.

Abstract

The aim of this doctoral thesis is the development of stochastic epidemic models focused on disease outbreaks in humans, as well as livestock. Statistical methodology is developed aimed at informing public health policies and their communication as implemented by the governing organizations, specifically at a time of crisis like the Covid-19 pandemic.

The second section is concerned with the results of a simulation-based evaluation of several policies for vaccine roll-out. Particular focus is placed upon on the effects of delaying the second dose of two-dose vaccines. In the presence of limited vaccine supply, the specific policy choice was a pressing issue for several countries worldwide, and the adopted course of action affected the extension or easing of non-pharmaceutical interventions (NPIs). We used a suitably generalised, age-structured, stochastic SEIR (Susceptible $\rightarrow$ Exposed $\rightarrow$ Infectious $\rightarrow$ Removed) epidemic model that accommodates quantitative descriptions of the major effects resulting from distinct vaccination strategies. The different rates of social contacts among distinct age-groups (as well as other model parameters) are informed by a

recent survey conducted in Greece, but the conclusions are widely applicable. The results are summarised and evaluated in terms of the total number of deaths and infections as well as life years lost.

A number of NPIs had been implemented in order to reduce transmission, thus leading to multiple phases of transmission. The disease reproduction number R_t, a way of quantifying transmissibility, has been a key part in assessing the impact of such interventions. In the third chapter of this thesis we discuss the distinct types of transmission models used and how they are linked. We consider a hierarchical stochastic epidemic model with piece-wise constant R_t, appropriate for modelling the distinct phases of the epidemic and quantifying the true disease magnitude. The location and scale of R_t changes are inferred directly from data while the number of transmissibility phases is allowed to vary. We determine the model complexity via appropriate Poisson point process and Dirichlet process-type modelling components. The models are evaluated using synthetic data sets and the methods are applied to freely available data from the United Kingdom and Greece as well as California and New York states. We estimate the true infected cases and the corresponding R_t, among other quantities, and independently validate the proposed approach using a large seroprevalence study.

Chapter four is concerned with a class of models where the Ornstein-Uhlenbeck (OU) process is embedded within Poisson-type point processes.

We utilise a general OU model with Student's t-distribution marginals and a Cox-Ingersoll-Ross model for the latent infection rate of the spatio-temporal model. We also propose a class of Bayesian Neural Nets using horseshoe priors for the weights. Real data from Foot and Mouth and Sheep-pox outbreaks in livestock within the Evros region of Greece are studied. The predictive ability of each model is being assessed using proper scoring rules within the prequential analysis framework. Our investigation concludes that the Student-t OU and the CIR models improve upon the previously introduced models with Gaussian OU for the latent rate of the Poisson-type point process.

Περίληψη

Στόχος της παρούσας διδακτορικής διατριβής είναι η ανάπτυξη στοχαστικών επιδημικών μοντέλων με έμφαση στις μολυσματικές ασθένειες σε ανθρώπους και ζώα. Αναπτύσσεται συγκεκριμένη στατιστική μεθοδολογία για να ενημερώνει καλύτερα τις δημόσιες πολιτικές υγείας και τις επικοινωνίες που υλοποιούνται από τις κυβερνητικές οργανώσεις, ειδικά κατά τη διάρκεια κρίσεων, όπως η πανδημία του Covid-19.

Η δεύτερη ενότητα αφορά τα αποτελέσματα μιας αξιολόγησης που βασίζεται σε προσομοίωση πολλών πολιτικών για την ανάπτυξη εμβολίων. Ιδιαίτερη έμφαση δίνεται στις επιπτώσεις της καθυστέρησης της δεύτερης δόσης των εμβολίων δύο δόσεων. Παρουσία περιορισμένης προμήθειας εμβολίων, η συγκεκριμένη επιλογή πολιτικής ήταν ένα πιεστικό ζήτημα για πολλές χώρες παγκοσμίως και η υιοθέτηση πορείας δράσης επηρέασε την επέκταση ή τη χαλάρωση των μη φαρμακευτικών παρεμβάσεων. Χρησιμοποιήσαμε ένα κατάλληλα γενικευμένο, ηλικιακά δομημένο, στοχαστικό μοντέλο επιδημίας **SEIR** (Ευάλωτοι $\rightarrow$ Εκτεθειμένοι $\rightarrow$ Μολυσματικοί $\rightarrow$ Αφαιρεθέντες) που περιλαμβάνει ποσοτικές περιγραφές των κύριων επιπτώσεων που προκύπτουν από διαφορετικές στρατηγικές εμβολιασμού. Τα διαφορετικά ποσοστά κοινωνικών επαφών μεταξύ διαφορετικών ηλικιακών ομάδων (καθώς και άλλες παραμέτρους του μοντέλου) ενημερώνονται από μια πρόσφατη έρευνα που διεξήχθη στην Ελλάδα, αλλά τα συμπεράσματα είναι ευρέως εφαρμόσιμα. Τα αποτελέσματα

συνοψίζονται και αξιολογούνται ως προς τον συνολικό αριθμό θανάτων και μολύνσεων καθώς και τα χαμένα χρόνια ζωής.

Στο τρίτο κεφάλαιο αυτής της διατριβής συζητάμε τους διαφορετικούς τύπους μοντέλων μετάδοσης που χρησιμοποιούνται και πώς συνδέονται. Θεωρούμε ένα ιεραρχικό στοχαστικό επιδημικό μοντέλο με αποσπασματική σταθερά R_t, κατάλληλο για τη μοντελοποίηση των διακριτών φάσεων της επιδημίας και τον ποσοτικό προσδιορισμό του πραγματικού μεγέθους της νόσου. Η θέση και η κλίμακα των αλλαγών R_t συνάγονται απευθείας από τα δεδομένα, ενώ ο αριθμός των φάσεων μεταδοτικότητας επιτρέπεται να ποικίλλει.Καθορίζουμε την πολυπλοκότητα του μοντέλου μέσω κατάλληλων διαδικασιών Poisson και Dirichlet. Τα μοντέλα αξιολογούνται χρησιμοποιώντας συνθετικά σύνολα δεδομένων και οι μέθοδοι εφαρμόζονται σε ελεύθερα διαθέσιμα δεδομένα από το Ηνωμένο Βασίλειο και την Ελλάδα καθώς και από τις πολιτείες της Καλιφόρνια και της Νέας Υόρκης. Υπολογίζουμε τα πραγματικά μολυσμένα κρούσματα και το αντίστοιχο R_t, μεταξύ άλλων ποσοτήτων, και επικυρώνουμε ανεξάρτητα την προτεινόμενη προσέγγιση χρησιμοποιώντας μια μεγάλη μελέτη οροεπιπολασμού.

Το τέταρτο κεφάλαιο ασχολείται με μια κατηγορία μοντέλων όπου η διαδικασία Ornstein-Uhlenbeck (OU) είναι ενσωματωμένη σε διαδικασίες τύπου Poisson. Χρησιμοποιούμε ένα γενικό μοντέλο OU με μεταβατική πυκνότητα πιθανότητας Student-t και ένα μοντέλο Cox-Ingersoll-Ross για το ποσοστό

μόλυνσης του χωροχρονικού μοντέλου. Προτείνουμε επίσης μια κατηγορία νευρωνικών δικτυών Bayes που χρησιμοποιούν εκ των προτέρων κατανομές τύπου horseshoe. Μελετώνται πραγματικά δεδομένα από επιδημίες αφθώδους πυρετού και ευλογιάς σε ζώα στην περιοχή του Έβρου στην Ελλάδα. Η προγνωστική ικανότητα κάθε μοντέλου αξιολογείται χρησιμοποιώντας τους κατάλληλους κανόνες βαθμολόγησης εντός του πλαισίου προκαταρκτικής ανάλυσης. Η έρευνά μας καταλήγει στο συμπέρασμα ότι τα μοντέλα Student-t OU και CIR βελτιώνουν τα μοντέλα που εισήχθησαν προηγουμένως με την Gaussian OU για τον ρυθμό της διαδικασίας Poisson.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Epidemiology is the study and analysis of the distribution and factors of health and disease conditions in defined populations. It is the foundation of public health and shapes policy decisions and evidence-based practice by identifying risk factors for disease and preventive healthcare. In the era of Covid19, it is evident more than ever that disease outbreak modelling is a crucial arrow in the quiver of health organizations worldwide. Epidemic modelling has flourished as a research area in the years before this global pandemic, thus enabling evidence-based decision-making. At the start of 2020, the emergence of the Covid-19 pandemic brought unprecedented challenges and conditions across the globe. In the absence of pharmaceutical options, nonpharmaceutical interventions (NPIs) have been implemented in order to reduce transmission. The reproduction number R_t, a way of quantifying transmissibility, has been

a key part in assessing the efficiency of these interventions. However, due to severe under-reporting and bias of reported cases, the true prevalence of the disease remains unknown.

Since the emergence of multiple vaccines, a gradual transition towards some form of 'normality' was initiated with the disease moving towards endemicity. A unified framework encapsulating the distinct phases of the pandemic in Greece and other European countries is represented in Chapter 2 of this thesis.

Standard epidemic theory (Andersson and Britton, 2000) suggests that in order to prevent major outbreaks, a proportion of at least $1 - \frac{1}{R}$ will have to become immune (either through vaccination or previous infection). In addition to social distancing (Lewnard and Lo, 2020) and mass testing (Taipale et al., 2021), the fair allocation of scarce medical interventions, such as vaccine distribution early on, presents ethical challenges as there are different allocation options like treating everyone equally, favouring the most vulnerable, maximising total benefit and no single principle can address all morally relevant considerations (Emanuel et al., 2020; Persad et al., 2009). Modelling studies broadly agree that when vaccine supply is limited, prioritising the most vulnerable (the elderly in the case of covid) is the optimal strategy to reduce COVID-19 mortality [10,11]. This is in agreement with epidemic theory (Andersson and Britton, 2000) which suggests that the focus for dis-

ease control should be based on a combination of targeting susceptibility and infectivity. In this work, we focus on the problem which many European countries were facing in the spring of 2021 whence the prioritisation of vaccines was of the essence. Due to supply constraints, it was decided in the UK and Canada to try and cover a larger fraction of the population by delaying the administration of the second dose since the first dose of the SARS-CoV-2 vaccine offers considerable protection (GOV.CA, 2021; GOV.UK, 2021).

In Chapter 2 we performed a simulation study using a novel stochastic age-structured compartmental model that accounts for distinct vaccination states. The model, termed S(V)EIR (Susceptible-Vaccinated-Infected-Removed) accounts for the age composition of the population, the social mixing rates of different age groups, the intention to get vaccinated as well as the appropriate risk of death for each age group. This model was used in order to evaluate different vaccination strategies and the associated reduction in Covid19-related mortality by delaying the timing of the second dose of the vaccine in order to administer the vaccine to a larger population fraction as early as possible. This work has been sent as a technical report to the Greek Covid19 response authorities and has been accepted for publication in PLOS ONE.

The reproduction number, R, of an epidemic, is a measure of the transmissibility of the disease and is of vital importance for informing the interventions to mitigate disease spread. In a covid-like outbreak, where interventions

affect disease transmissibility, it is advantageous to estimate the instantaneous reproduction number R_t. A wide range of methods have been proposed for estimating R_t from disease surveillance data. The Wallinga and Teunis method (Wallinga and Teunis, 2004) requires only case incidence data and the distribution of the serial interval (the time between the onset of symptoms in a primary case and the onset of symptoms of secondary cases) for estimating R_t. The drawback of this method is that in order to calculate Rt at day t one needs data beyond day t. The authors in Cori et al. (2013) amended this technique using branching processes building on the work of Fraser (2007) where the incidence on day t is calculated as the weighted average of the incidences on the previous days multiplied by the instantaneous reproduction number R_t. The authors further extend this methodology by using a hierarchical Bayesian model which allowed them to work with the reported death cases and estimate the true disease prevalence. A detailed presentation of the compartmental stochastic epidemic model is given in both chapters 2 and 3. Also, Chapter 3 contains a thorough presentation of the equivalence of the stochastic Susceptible-Infected-Recovered model with the renewal process epidemic models.

Statistical inference for the models entertained in this thesis is mostly performed under the Bayesian paradigm. An inherent problem with epidemic models is that the required data are never fully observed. This issue is further

exacerbated in the case of SARS-CoV2 where a large proportion of the population remains asymptomatic and there is under-ascertainment of the true number of infections. In Chapter 3, We opt to use a Bayesian hierarchical model, where the source of information is the daily reported deaths which are likely less prone to under-reporting. We build upon the work of Flaxman et al. (2020) and amend their methodology by using a Dirichlet process and a Poisson process prior (Blackwell and MacQueen, 1973; Ishwaran and Zarepour, 2002) to facilitate the semi-automatic determination of the number of transmission waves. We infer the position and the magnitude of changes in the transmission rate and predict the phases of the epidemic directly from the data. Alternative constructions of the Dirichlet process are employed and compared in terms of statistical accuracy and computational efficiency. The complexity of the developed models necessitates the use of state-of-the-art learning techniques and to this end, we use Hamiltonian Monte Carlo and related methods implemented in R and Python programming languages.

Epidemic models for animal diseases such as foot and mouth (FMD) or sheep-pox are typically used in order to characterise disease transmission and inform the decision-making process of the relevant organizations. The highly contagious nature of such viruses results in dire consequences on the animals' well-being and significant economic consequences for the professionals involved in animal husbandry. Therefore, it is crucial for epidemiological

models to accurately predict the course of outbreaks and mitigate disease spread.

In Chapter 4 our work on livestock epidemics builds upon the work of Malesios et al. (2017). The authors used a log Gaussian Cox process (Møller et al., 1998) to perform inference and variable selection on the number of infected farms in Evros, Greece. In particular, they used an Ornstein-Uhlenbek (OU) process (Uhlenbeck and Ornstein, 1930) embedded inside a Poisson process. The use of Gaussian processes is popular in the relevant literature (Diggle et al., 2013), especially in the presence of temporal dependence in the data. The authors extended the model to account for the spatial nature of the data using appropriate kernel functions. Variable selection was performed using the methodology presented in Dellaportas et al. (2002). We further extend these models with the introduction of OU-type processes with Student's t-distribution transition densities and the Cox-Ingersoll-Ross type model for the latent rate of the Cox process. These models are compared against Bayesian non-parametric models involving Bayesian artificial neural networks using prequential analysis (Dawid, 1984).

1.2 Bayesian Methodology

1.2.1 Bayesian Inference

In this section, we will present at a non-technical level the basic concepts of the Bayesian methodology. For a more detailed analysis, we direct the reader to the book of Bernardo and Smith (1994).

Under the Bayesian paradigm, in addition to defying a model for the observed data $y = (y_1, y_2, ..., y_n)$ in the form of the likelihood function $L(y|\theta)$ given the vector of the random parameters θ, we also define the prior distributions of these parameters $\pi(\theta)$. The prior distributions can demonstrate previous knowledge about these parameters or a complete lack thereof (non-informative priors). Then inference about θ is performed through the Bayes' theorem and their posterior distribution:

$$\pi(\theta|y) = \frac{L(y|\theta)\pi(\theta)}{\int L(y|\theta)\pi(\theta)} \tag{1.1}$$

The appeal of sampling from the posterior distribution lies in the fact that we can estimate any desired statistic of a posterior distribution by ergodic averages, given that we have N samples from that distribution. Hence for every function of the parameter of interest $G(\theta)$, we can calculate the posterior mean for example by simply:

- Generating a sample $\theta_1, \theta_2, ..., \theta_T$ from the posterior distribution $\pi(\theta|y)$.

- Calculate the sample mean of $G(\theta)$ by simply calculating the quantity:

$$\frac{1}{T}\sum_{i=1}^{T} G(\theta_i) \tag{1.2}$$

The main problem in the above-mentioned procedure is how to generate from the posterior density $\pi(\theta|y)$. In most cases, it is not straightforward, since the integral in the denominator is not analytically available in all but the most simple cases, where there is conjugacy between the prior and the likelihood functions. The generality and flexibility of Markov chain-based simulation techniques can overcome these difficulties. We will present some generic Markov Chain Monte Carlo (MCMC) algorithms that have been pivotal in the emergence and development of Bayesian inference.

1.2.2 Bayesian Computation

When the posterior distribution is not analytically available, MCMC algorithms are used to construct a Markov chain, which has the desired posterior distribution as its equilibrium. By simulating the Markov chain for a sufficient amount of time, the samples generated eventually converge to samples from the posterior distribution. It dates back to the pioneering paper of Metropolis et al. (1953) although there was a lack of computational power available at the time. The generalization of the sampling method was proposed by Hastings (1970) in the Metropolis-Hastings algorithm that is presented below.

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm generates samples from a target density $\pi(\theta)$, which is only known up to a normalizing constant. The samples are generated through a user-selected and calibrated proposal function $q(\theta)$ and are accepted or rejected after a Metropolis acceptance step. The proper selection of the proposal ensures the irreducibility and aperiodicity of the Markov chain and as a result the sampling of the correct posterior distribution. The selection of the proposal distribution also affects the convergence rate of the Markov chain to the target distribution. The algorithm is shown below:

Algorithm 1 Metropolis-Hastings Algorithm

1: Start with an arbitrary initial value θ_0

2: **for** $n = 0, 1, 2, \ldots$ **do**

3: Generate $\xi \sim q(\xi|\theta_n)$

4: Evaluate $\alpha = \min\left(1, \frac{\pi(\xi)q(\theta_n|\xi)}{\pi(\theta_n)q(\xi|\theta_n)}\right)$

5: Set

6: $$\theta_{n+1} = \begin{cases} \xi, & \text{with probability } \alpha, \\ \theta_n, & \text{otherwise.} \end{cases}$$

7: **end for**

Gibbs sampling

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm. When the conditional densities of each parameter are analytically available, we can sample sequentially each parameter conditional to the rest of the

parameters and acquire a sample from the target distribution. In real-world problems, the full conditionals may not be available or tractable. These steps can be substituted by a Metropolis-Hastings step. These mixtures of Gibbs and Metropolis-Hastings samplers are used by widely available programming languages like WinBUGS, JAGS or the R package nimble. The Gibbs sampler algorithm is shown below:

Algorithm 2 Gibbs Sampling Algorithm

Input: θ_0, M
Set $\theta \leftarrow \theta_0$
for $m = 1$ **to** M **do**
    for $i = 1$ **to** n **do**
        Sample $\theta_i \sim p(\theta_i|\theta_{-i})$
    end for
    Set $\theta \leftarrow \theta_1, \theta_2, \ldots, \theta_n$
end for

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC), also known as Hybrid Monte Carlo, is an MCMC method that uses the spatial characteristics of the target distribution, through its derivatives, in order to generate efficient transitions and samples spanning the space of the posterior (Betancourt and Girolami, 2013; Neal, 2011). It uses numerical integration on an approximate Hamiltonian dynamics system, which is then corrected using a Metropolis acceptance step. HMC

contrary to the Metropolis-Hastings and Gibbs algorithms avoids the random walk behaviour when exploring the posterior distribution. In order for HMC to generate an independent sample from a target distribution of dimension D is approximately $O(D^{\frac{5}{4}})$, which stands in sharp contrast with the $O(D^2)$ cost of random-walk Metropolis (Hoffman and Gelman, 2014). The main drawback of HMC is the high computational cost of each iteration since the Hamiltonian dynamics must be calculated, which requires the derivatives of the target distribution up to a normalizing constant. These derivatives in many applications-models can be cumbersome to calculate or even impossible. This problem is facilitated through the use of automatic differentiation.

HMC introduces auxiliary momentum variables ρ and together with the position variables θ, the parameters of the model, they define the joint posterior density:

$$p(\rho,\theta|y) = p(\rho|\theta,y)p(\theta|y) \tag{1.3}$$

The joint density $p(\rho,\theta|y)$ defines a Hamiltonian:

$$H(\rho,\theta|y) = -\log(\rho,\theta|y) \tag{1.4}$$

$$\begin{aligned} H(\rho,\theta|y) =& -\log(\rho,\theta|y) \\ =& -\log p(\rho|\theta,y) - \log p(\theta|y) \\ =& T(\rho|\theta,y) + V(\theta|y) \end{aligned}$$

where the term $T(\rho|\theta,y) = -\log p(\rho|\theta,y)$ is called kinetic energy and the term $V(\theta|y) = -\log p(\theta|y)$ is called potential energy. The potential energy under the Bayesian paradigm is the log-likelihood plus the logarithm of the prior of the parameters θ.

Starting from the current value of the parameters, a transition to a new state is generated in two stages before being subjected to a Metropolis acceptance step. First, a value for the momentum is drawn independently of the current parameter values.

$$\rho \sim \mathrm{N}(0,\Sigma) \tag{1.5}$$

Thus momentum does not persist across iterations. Next, the joint system $p(\rho,\theta|y)$ made up of the current parameter values θ and new momentum ρ is evolved via Hamilton's equations.

$$\begin{aligned}\frac{d\theta}{dt} &= +\frac{\partial H}{\partial \rho} = +\frac{\partial T}{\partial \rho} \\ \frac{d\rho}{dt} &= -\frac{\partial H}{\partial \theta} = \frac{\partial V}{\partial \theta}\end{aligned}$$

With the momentum density being independent of the target density and the data, $p(\rho|\theta,y) = p(\rho)$, the first term in the momentum time derivative $\frac{\partial T}{\partial \theta}$ is zero.

This two-state differential equation is solved by using the leapfrog integrator. The leapfrog integrator is a numerical integration algorithm that

is specifically adapted to provide stable results for Hamiltonian system of equations. The leapfrog integrator takes discrete steps of some small time integral ε. It begins by drawing a momentum value from the ρ density and the alternate half-step updates of the momentum and full-step updates of the position.

$$\begin{aligned}\rho &\leftarrow \rho - \frac{\varepsilon}{2}\frac{\partial V}{\partial \theta} \\ \theta &\leftarrow \theta + \varepsilon\rho \\ \rho &\leftarrow \rho - \frac{\varepsilon}{2}\frac{\partial V}{\partial \theta}\end{aligned}$$

After L leapfrog steps are applied, a total of $L\varepsilon$ time is simulated. At the end, in order to account for numerical errors during integration, we apply a Metropolis acceptance step. The complete HMC algorithm is shown below:

Algorithm 3 Hamiltonian Monte Carlo

Input: θ_0, ε, L, L, M
for $m = 1$ to M **do**
    Sample $\rho_0 \sim N(0, I)$
    Set $\theta_m \leftarrow \theta_{m-1}$, $\tilde{\theta} \leftarrow \theta_{m-1}$, $\tilde{\rho} \leftarrow \rho_0$
    for $i = 1$ to L **do**
        Set $\tilde{\theta}, \tilde{\rho} \leftarrow \text{Leapfrog}(\tilde{\theta}, \tilde{\rho}, \varepsilon)$
    end for
    Compute acceptance probability $\alpha = \min\left(1, \frac{\exp\left(L(\tilde{\theta}) - \frac{1}{2}\tilde{\rho}\cdot\tilde{\rho}\right)}{\exp\left(L(\theta_{m-1}) - \frac{1}{2}\rho_0\cdot\rho_0\right)}\right)$
    if $\alpha \geq \text{Uniform}(0,1)$ **then**
        Set $\theta_m \leftarrow \tilde{\theta}$, $\rho_m \leftarrow -\tilde{\rho}$
    end if
end for

Algorithm 4 Leapfrog

function $\text{Leapfrog}(\theta, \rho, \varepsilon)$
    Set $\tilde{\rho} \leftarrow \rho + \left(\frac{\varepsilon}{2}\right)\nabla_\theta L(\theta)$
    Set $\tilde{\theta} \leftarrow \theta + \varepsilon\tilde{\rho}$
    Set $\tilde{\rho} \leftarrow \tilde{\rho} + \left(\frac{\varepsilon}{2}\right)\nabla_\theta L(\tilde{\theta})$
    return $\tilde{\theta}, \tilde{\rho}$
end function

No-U-Turn Sampler

HMC's increased efficiency in sampling comes at an increased cost. HMC requires that the user specify the step size ε and the number of steps L for which to run a simulated Hamiltonian system. The selection of these

parameters is crucial in order to reap the benefits of the HMC but in reality, they are difficult to calibrate. A poor choice of either of these parameters will result in a dramatic drop in HMC's efficiency. Proper calibration requires many trial runs increasing the computational cost, as well as deep knowledge and understanding of the algorithm in order to interpret the results of these trial runs. If ε is too large, the simulation will be inaccurate, on the other hand, if ε is too small, the computation will be wasted taking many small steps. If L is too small, successive samples will be too close resulting in random walk behaviour and slow mixing. If L is too large, then HMC will generate trajectories that loop back and retrace the steps. The No-U turn sampler (NUTS) was introduced, an MCMC sampler that retains HMC's ability to overcome random walk behaviour and eliminates the need to tune these parameters making the HMC available for even the uninitiated users.

NUTS starts by introducing a slice variable u with conditional distribution $p(u|\theta,\rho) = \text{Uniform}([0, \exp L(\theta_{m-1}) - \frac{1}{2}\rho\cdot\rho])$, which renders the conditional distribution of θ and ρ given u Uniform. After sampling $u|\theta,\rho$ NUTS uses the leapfrog integrator to simulate trajectories forward or backwards in time, first 1 step forward or backwards, then 2 steps forwards or backwards, then 4 steps etc. The resulting binary tree has leaf nodes corresponding to momentum-position states. The process stops when the sub trajectory from the leftmost to the rightmost nodes of any balanced subtree of the overall

binary tree starts to backtrack on itself (i.e., makes a "U-turn"). Then a state from the ones generated is randomly chosen by using a transition kernel that leaves invariant the uniform distribution over the set of all the states we can transition without violating the balance (i.e. their joint probability is above the slice variable u). By running the Hamiltonian simulation both forward and backward in time NUTS ensures the reversibility of the Markov chain and guarantees convergence to the target distribution. The NUTS algorithm was used in this thesis through its implementation in probabilistic programming language *Stan* (Stan Development Team, 2023). The complete algorithm is shown below:

Algorithm 5 No-U-Turn Sampler (NUTS)

Input: $\theta_0, \varepsilon, L, M$

for $m = 1$ **to** M **do**

  Resample $\rho_0 \sim N(0, I)$

  Resample $u \sim \text{Uniform}([0, \exp L(\theta_{m-1}) - \frac{1}{2}\rho_0 \cdot \rho_0])$

  Initialize $\theta_- = \theta_{m-1}, \theta_+ = \theta_{m-1}, \rho_- = \rho_0, \rho_+ = \rho_0, j = 0, \theta_m = \theta_{m-1}, n = 1, s = 1$

  while $s = 1$ **do**

    Choose a direction $v_j \sim \text{Uniform}(-1, 1)$

    if $v_j = -1$ **then**

      $\theta_-, \rho_-, -, -, \theta_0, n_0, s_0 \leftarrow \text{BuildTree}(\theta_-, \rho_-, u, v_j, j, \varepsilon)$

    else

      $-, -, \theta_+, \rho_+, \theta_0, n_0, s_0 \leftarrow \text{BuildTree}(\theta_+, \rho_+, u, v_j, j, \varepsilon)$

    end if

    if $s_0 = 1$ **then**

      With probability $\min 1, \frac{n_0}{n}$, set $\theta_m \leftarrow \theta_0$

    end if

    $n \leftarrow n + n_0$

    $s \leftarrow s_0 \times \mathbb{I}[(\theta_+ - \theta_-) \cdot \rho_- \geq 0] \times \mathbb{I}[(\theta_+ - \theta_-) \cdot \rho_+ \geq 0]$

    $j \leftarrow j + 1$

  end while

end for

Algorithm 6 BuildTree Function

function BuildTree($\theta, \rho, u, v, j, \varepsilon$)
  if $j = 0$ **then**
    Base case - take one leapfrog step in the direction v
    $\theta_0, \rho_0 \leftarrow \text{Leapfrog}(\theta, \rho, v, \varepsilon)$
    $n_0 \leftarrow \mathbb{I}[u \leq \exp L(\theta_0) - \frac{1}{2}\rho_0 \cdot \rho_0]$
    $s_0 \leftarrow \mathbb{I}[L(\theta_0) - \frac{1}{2}\rho_0 \cdot \rho_0 > \log u - \Delta_{\max}]$
    return $\theta_0, \rho_0, \theta_0, \rho_0, \theta_0, n_0, s_0$
  else
    Recursion - implicitly build the left and right subtrees
    $\theta_-, \rho_-, \theta_+, \rho_+, \theta_0, n_0, s_0 \leftarrow \text{BuildTree}(\theta, \rho, u, v, j-1, \varepsilon)$
    if $s_0 = 1$ **then**
      if $v = -1$ **then**
        $\theta_-, \rho_-, -, -, \theta_{00}, n_{00}, s_{00} \leftarrow \text{BuildTree}(\theta_-, \rho_-, u, v, j-1, \varepsilon)$
      else
        $-, -, \theta_+, \rho_+, \theta_{00}, n_{00}, s_{00} \leftarrow \text{BuildTree}(\theta_+, \rho_+, u, v, j-1, \varepsilon)$
      end if
      With probability $\frac{n_{00}}{n_0 + n_{00}}$, set $\theta_0 \leftarrow \theta_{00}$
      $s_0 \leftarrow s_{00} \times \mathbb{I}[(\theta_+ - \theta_-) \cdot \rho_- \geq 0] \times \mathbb{I}[(\theta_+ - \theta_-) \cdot \rho_+ \geq 0]$
      $n_0 \leftarrow n_0 + n_{00}$
    end if
    return $\theta_-, \rho_-, \theta_+, \rho_+, \theta_0, n_0, s_0$
  end if
end function

Adaptively Tuning the step size ε

Having shown how to automatically set the number of steps L in HMC using the NUTS algorithm, we will now focus on the step size parameter e. This is performed by using stochastic optimization using a statistic H_t that describes the behaviour of an MCMC algorithm at iteration t. Usually, a function of the acceptance probability of the MCMC is used as H_t but the NUTS algorithm does not have an accept-reject step. For each iteration we define the statistic H_t^{NUTS} and its expectation when the chain has reached its equilibrium:

$$H_t^{NUTS} \equiv \frac{1}{B_t^{final}} \sum_{\theta,\rho \in B_t^{final}} \min\left\{1, \frac{p(\theta,\rho)}{p(\theta^{t-1},\rho^{t,0})}\right\}; \quad h^{NUTS} \equiv E[H_t^{NUTS}] \tag{1.6}$$

where B_t^{final} is the set of all states explored in the final doubling of the binary tree and $\theta^{t-1}, \rho^{t,0}$ are the initial position and resampled momentum of the tth iteration. H^{NUTS} can be understood as the average acceptance probability that HMC would give to the position-momentum states explored during the final doubling iteration of the NUTS algorithm. We set $H_t \equiv \delta - H^{NUTS}$ where δ is a pre-specified desirable value we wish to achieve.

The *STAN* programming language that implements the NUTS uses the dual averaging scheme of Nesterov (2007), an algorithm for non-smooth and stochastic convex optimization to update the step size automatically.

Assuming that we want to find a setting of a parameter $x \in \mathbb{R}$ such that $h(x) \equiv E_t[H_t|x] = 0$, we can apply the updates:

$$x_{t+1} \leftarrow \mu - \frac{\sqrt{t}}{\gamma}\frac{1}{t+t_0}; \quad \bar{x}_{t+1} \leftarrow \eta_t x_{t+1} + (1-\eta_t)\bar{x}_t \tag{1.7}$$

where μ is a freely chosen point that the iterates x_t are shrunk towards,$\gamma > 0$ is a free parameter that controls the amount of shrinkage towards μ, $t_0 \geq 0$ is a free parameter that stabilizes the initial iterations of the algorithm, $\eta_t \equiv t^{-k}$ and we define $\bar{x}_1 = x_1$ and $x \equiv \log\varepsilon$. In *STAN* the user only defines the value of the parameter δ and the step size is automatically adjusted.

Chapter 2

Evaluating the effects of second-dose vaccine-delay policies

2.1 Introduction

Since December 2019, COVID-19 has presented a global threat to public health and to the worldwide economy, and it will likely continue to disrupt livelihoods until a high percentage of the population is vaccinated. High vaccination rates will be necessary to reach herd immunity in a short period of time. Standard theory (Andersson and Britton, 2000) suggests that a proportion approximately equal to $1 - \frac{1}{R_0}$ of the population will have to become immune (either through vaccination or previous infection) in order to effectively suppress disease transmission, where R_0 is the virus' basic reproduction number. The actual vaccination coverage required is likely to vary due to population heterogeneity, previous levels of the spread of

infection, and other local factors. In addition, the exact value of R_0 for SARS-CoV-2 under "normal" conditions remains quite uncertain since there has been very little disease spread without some mitigation effort due to non-pharmaceutical-interventions (NPIs), and also due to the appearance of new variants. Therefore, constrained scenarios are likely to give a realistic estimate of the effect of distinct vaccination policies and this approach is adopted in the present paper.

Assuming a vaccination coverage between 60%-80% of the population, 3.1-4.1 billion people worldwide will need to be vaccinated (Wang et al., 2020). With several seemingly highly efficacious vaccines available (efficacy estimated at 94.1%, 95% and 62% for Moderna, Pfizer-BioNTech and Oxford-AstraZeneca respectively) against COVID-19 disease (Baden et al., 2021; FDA, 2020; Polack et al., 2020; Skowronski and De Serres, 2021; Voysey et al., 2020) it appears that a return to near-normality for society and for the economy may soon be possible. Unfortunately, limited supply is currently an impediment to achieving high vaccination coverage rapidly (Bollyky, 2021).

In addition to social distancing (Lewnard and Lo, 2020) and mass testing (Taipale et al., 2021), the fair allocation of scarce medical interventions such as vaccines presents ethical challenges as there are different allocation principles – treating people equally, favouring the worst-off, maximising total benefits, and promoting and rewarding social usefulness – and no single principle can

address all morally relevant considerations (Emanuel et al., 2020; Persad et al., 2009). Modelling studies broadly agree that when vaccine supply is limited, prioritising the elderly is a necessary strategy to reduce COVID-19 mortality, whereas the prioritisation of younger individuals would have an impact on reducing transmission (Bubar et al., 2021; Matrajt et al., 2021a). This agrees with epidemic theory (Andersson and Britton, 2000) which suggests that the focus for disease control should be based on a combination of targeting susceptibility and infectivity. Therefore, assuming very scarce resources, it makes sense to focus on the most vulnerable individuals in the population. On the other extreme is the presence of a nearly unlimited vaccine supply, whence aiming for achieving herd immunity is straightforward. In this work, we focus on the intermediate problem that many European countries are currently facing, and the prioritisation of vaccines is of the essence.

Due to supply constraints, it was decided in the UK and Canada to delay the administration of the second dose of all vaccines, based on the rationale that SARS-CoV-2 vaccination offers considerable protection after the first dose and that more people could benefit (GOV.CA, 2021; GOV.UK, 2021). Although this approach seems appealing, the impact of delaying the second dose is not straightforward as it depends on several parameters such as the efficacy of the first dose in time, the levels of transmission in the population, vaccination rollout, and the vaccine profile (reduction in symptoms or

in symptoms and infection) (Matrajt et al., 2021b; Paltiel et al., 2021a,b). Country-specific information on the age distribution of the population and social mixing patterns are also necessary to obtain realistic estimates.

The main contribution of this work is the evaluation of different vaccination strategies and their potential benefits, primarily based on data from Greece, a typical country of the EU area in terms of vaccine availability and administration, with a population of around 10.8 million people (ELSTAT, 2021). The current strategy (strategy I) is to give the second vaccine dose three weeks after the first for the Pfizer vaccine, which currently consists of the largest portion of the available vaccines in the EU. We consider an alternative policy (strategy II) where, after the vaccination of medical personnel and those over 75, a portion of the available vaccines is distributed with a three-month time interval between the two doses. The prioritization of the medical personnel and those over 75 years old is kept constant for every strategy. Our methodology examines scenarios where the two different vaccination schedules are combined in different proportions, allowing us to explore the optimal portion of the population that should be vaccinated using the extended three-month time interval between the two doses. This is something that has not been extensively explored in the literature since earlier studies primarily focus on finding the optimal timing of the second dose, considering that the entire population will follow the same schedule

(Ferreira et al., 2021; Moore et al., 2021; Silva et al., 2021). Extending the time interval between doses to three months aims for faster partial coverage of economically active individuals, therefore offering indirect protection to a larger proportion of the population and ultimately potentially reducing the pandemic cost to public health and society. This is implicitly informed by aiming for a combined effect of reducing susceptibility and infectivity in the population. Different scenarios of vaccine availability and transmission rates are considered, as well as different scenarios for the acquired immunity after the first dose for strategy II. We assess our results through simulation of an age-structured stochastic SEIR (Susceptible $\rightarrow$ Exposed $\rightarrow$ Infectious $\rightarrow$ Removed) epidemic model, suitably modified to account for the number of vaccinated individuals with different protocols. We opted for a stochastic model because the most effective way to describe the spread of a disease is stochastic, based on the specification of the probability of disease transmission between two individuals. One may incorporate additional sources of stochasticity in the length of latent and infectious periods, but it is well known that such uncertainty is immaterial in terms of its effect on the outcome of an epidemic, particularly in a large population setting such as country-level studies; see for example Diekmann et al. (2013). Since the writing of this paper, related modelling techniques using deterministic dynamics have been proposed, that use optimization techniques to infer the model parameters and

find the optimal dosing schedule (Ferreira et al., 2021; Silva et al., 2021). An extension of a discrete-time, deterministic susceptible-infected-recovered model was used in Parino et al. (2021) to plan the scheduling of first and second vaccine doses, with the underlying objective of the optimization problem being the concurrent minimization of both the healthcare impact of the epidemic and of the socio-economic impact due to the implementation of NPIs. A similarly extended SEIR model was also used to measure (through simulation) whether the effect of a standard vaccination schedule for different uptake scenarios is enough to stop the epidemic without the need of NPIs (Moore et al., 2021).

2.2 Materials and methods

2.2.1 The multitype S(V)EIR model and simulation description

The model used for the simulation of different vaccination strategies is an age-structured stochastic SEIR model that accounts for different vaccinated populations and vaccinated states, termed S(V)EIR henceforth. A schematic representation of the model is given in Figure 2.1. In order to evaluate the effects of different vaccination strategies, this model also accounts for the age composition of the population, the social mixing rates of different age groups, the intention to get vaccinated, as well as the different risk of death of each age group.

The code for simulating the model is publicly available at: https://github.com/pbarmpounakis/Evaluating-the-effects-of-vaccine-rollout-policies-in-European-countries-A-simulation-stud

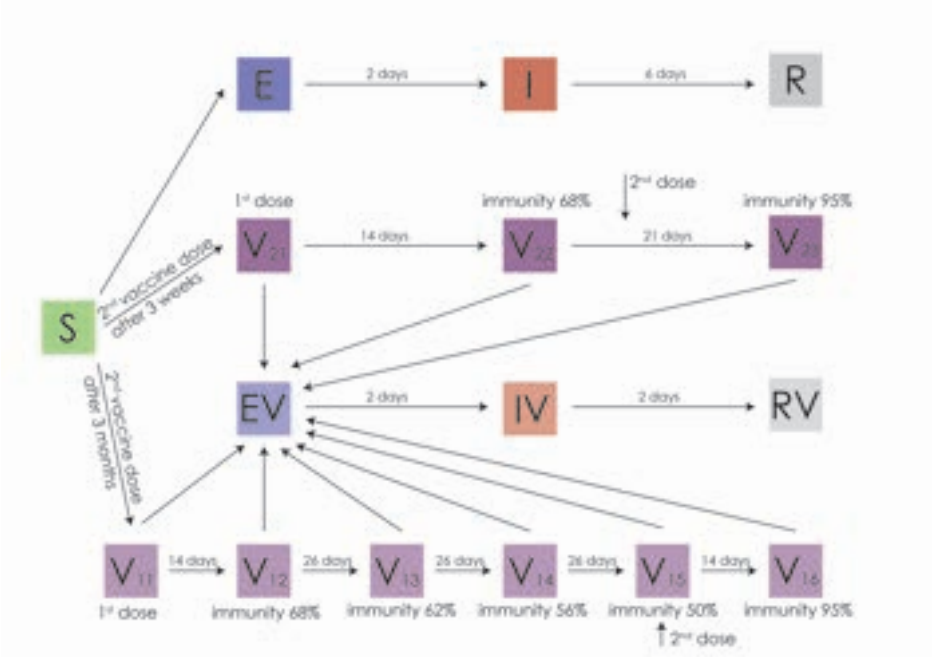


Figure 2.1 Schematic representation of the S(V)EIR epidemic model for the baseline scenario of immunity waning.

A detailed description of the model follows, while a summary of the quantitative assumptions made is given in Section 2.4.

States and vaccination effect assumptions

Two groups are considered for the vaccinated people representing the two distinct vaccination categories. In vaccination group 1, individuals receive the 2^{nd} dose of the vaccine after 3 months while in vaccination group 2 it is given after 3 weeks. In both vaccination groups, individuals who received the 1^{st} dose of the vaccine move to states V11 and V21 respectively and

remain fully susceptible. Two weeks after the 1st dose individuals from both vaccination groups move to the second stage (V12, V22 respectively), whence immunity jumps to 68% (Voysey et al., 2020). Individuals from vaccination group 2 remain at V22 for another 7 days when they take their 2nd dose and move to state 3 (V23) with their immunity jumping at 95% after two weeks. Individuals from vaccination group 1 take their second dose 78 days after entering V12 and then move to state V15 with their immunity changing based on different waning immunity scenarios described below. The stages V13, V14 and V15 account for the drop of immunity due to waning vaccine efficacy 26, 52 and 78 days after the first dose, respectively. They move to V16 14 days later when their immunity jumps to 95%, pertaining to the reported efficacy of the mRNA-based vaccines which are mostly used in European countries (Baden et al., 2021; FDA, 2020; Polack et al., 2020; Skowronski and De Serres, 2021; Voysey et al., 2020).

Transmission model assumptions

New infections from each state s and age group i follow a Binomial distribution with size given by the number of people in state s and age group i, and infection probability $(1 - immunity_s) * (1 - e^{-\sum_{j=1}^{n\_groups} \lambda_{ij} * \frac{I_j}{N_j}})$, where $immunity_s$ is the level of immunity at stage s, I_j is the number of infectious individuals at age group j, N_j is the total number of individuals at age group j, λ_{ij} is the i, j entry in the transmission matrix λ, and n_{groups} is the total number of different

age groups(Andersson and Britton, 2000). Following infection, individuals of age group i follow the Exposed $(E_i) \rightarrow Infectious(I_i) \rightarrow Removed(R_i)$ path with a constant exposure time of 2 days, based on an average incubation time of approximately 5 days (Lauer et al., 2020; Li et al., 2020b; Liu et al., 2020) and assuming that infectiousness starts approximately 2 days prior to the occurrence of symptoms (Ganyani et al., 2020; He et al., 2020; Li et al., 2020a). The infectious period is also assumed constant and set at 6 days for non-vaccinated individuals (Bi et al., 2020; Cereda et al., 2021; Lavezzo et al., 2020) and 2 or 3 days for vaccinated ones, depending upon the scenario of immunity waning and vaccine efficacy in reducing the infectious period (de Gier et al., 2021; Singanayagam et al., 2022). The choices of these values are conservative assuming that in reality people experiencing influenza-like symptoms will get tested and self-isolate, resulting in a lower effective infectious period. The total number of deaths is computed by multiplying the number in R_i with the infection fatality ratio (IFR) of each age group for the unvaccinated individuals as it was reported by CDC (2020) and with IFR x 5% for those vaccinated (Haas et al., 2021).

Different scenarios for R_t and immunity

Transmission levels corresponding to a constant effective reproduction number $R_t = 1.2$ and $R_t = 1.4$ are considered along with various levels of immunity at each stage of vaccination for group 1. These choices of R_t suggest

moderate transmission levels without the presence of a 'hard lockdown' for extended periods of time, closely resembling the policy that most European counties implement regarding social distancing measures during the vaccination period. R_t is calculated as the largest eigenvalue of the next-generation matrix, using an appropriate contact matrix. The next generation matrix G has elements $g_{ij} = \lambda_{ij}E(I) = \xi c_{ij}E(I)$, where $E(I)$ is the average period an individual remains infectious for, ξ is the chance of getting infected upon contact and c_{ij} is the i,j element of contact matrix C. R_t is set to a specific value by changing ξ. The contact matrices used for the calculations of the values of R_t were based on a social contacts survey assessing contacts in Greece; we have used the data collected in the second half of September 2020. This contact matrix informs the relative infectivity between age groups but, importantly, the scale is set by the value of R_t. We consider reduced infection probability for children by 48% (Haas et al., 2021).

We ran 1000 simulations for each scenario and computed the median as well as 90% equal-tailed uncertainty intervals. As precise data are not available for the precise course of infectivity and acquired immunity, three scenarios are considered.

Worst Case Scenario: It is assumed that, during the three months between the first and second dose (strategy II), the acquired immunity drops linearly

to 34% (Figure 2.5). The effective infectious period of those vaccinated is reduced by 50% to 3 days.

Baseline Scenario: Here it is assumed that during the three months between the two doses (strategy II) the acquired immunity drops linearly to 50% and the infectious period of those vaccinated is set at 2 days (Fig 1).

Optimistic Scenario: In this case, a constant immunity of 68% is assumed for the entire time between the first and second dose, and the infectious period lasts 2 days (Figure 2.4)).

Fraction of vaccines given to general population

Different percentages are considered for the proportion of the available vaccines distributed under strategy II. These are set to 0% (strategy I), 20%, 50% and 100% (strategy II); the resulting number of deaths and life years lost are computed in each of these cases. Moreover, in the model we assume that individuals due for the second dose have priority over those waiting to have their first dose, keeping the time interval between the two doses intact. The remaining doses available each day are given to unvaccinated individuals.

Initial conditions

We assume that the number of susceptibles at the start of the simulations is $S_{initial} = [N - R_{initial} - I_{initial}]$, where $R_{initial}$ is the estimated number of people

having gotten infected and recovered or died ∼620000 and N = 10816287, the total population of Greece. The active infectious population $I_{initial}$ is assumed ∼60000. These assumptions are informed by using the total number of deaths in Greece for each age group and the corresponding infection fatality ratio (IFR) in November of 2020, around 6 months after the first confirmed Covid-19 related death. These conditions are the same for all the scenarios for R_t and immunity.

Vaccine availability

Two levels of vaccine availability are considered; a baseline level and a limited level with a reduced number of vaccines, see Table 2.6.

Intention to get vaccinated

The populations' intention to get vaccinated is informed by a telephone survey contacted by Sypsa et al. (2021a); see Table 2.7. We assume that after the vaccination coverage of an age group reaches the percentage of people answering 'Probably/Definitely Yes' to whether they intend to get vaccinated, the vaccination rollout continues to the next (younger) age group. The percentage of people answering 'Probably/Definitely No' remains unvaccinated. The individuals answering 'Don't know/Don't answer' are distributed among the two other groups according to their respective adjusted percentages.

2.3 Results

Our main finding is that the optimal strategy in terms of the reduction in cumulative number of deaths (Figure 2.2) and number of years lost, is the one where all available vaccine doses are given under strategy II, using a time interval of three months between the two doses (Table 2.1).

The results vary between different immunity waning scenarios and different values of R_t, but they are robust in that the optimal strategy is always found to be the one that allocates 100% of the available doses under strategy II. The intermediate allocations of 20% or 50% of the available doses for strategy II show similar mortality for the whole population with Strategy I. But because fewer younger people get infected and die, there is a reduction in total life years lost (see Section 2.4). The total number of years of life lost is computed by summing over the different age-groups the corresponding $d\mu$ product, where d denotes the number of deaths in each age-group and μ is the difference between the total life expectancy in Greece (82 years) and the average age of each age group.

Next, we examine the resulting figures for cumulative number of deaths, daily deaths, life years lost, and daily infections, under the baseline immunity waning scenario, with $R_t = 1.2$, and with standard vaccine availability (Figures 2.2,2.3, 2.6, 2.7); detailed summaries of simulation results are presented in Section 2.4; several additional results based on both the optimistic and

worst-case scenarios are summarised in a web supplement at:

https://github.com/pbarmpounakis/Evaluating-the-effects-of-vaccine-rollout-policies-in-European-countries-A

Table 2.1 Comparison of strategy I and strategy II (at 100% doses given) for 2021 with $R_t = 1.2$, under the baseline immunity scenario and standard vaccine availability. "Gain" refers to the number of fewer deaths and life years lost under strategy II (extended interval between doses).)

	Strategy II vs. strategy I
Total (%) reduction under strategy II during January-December 2021	
Gain in number of deaths	579 (9.04%)
Gain in years of life	14802 (10.65%)

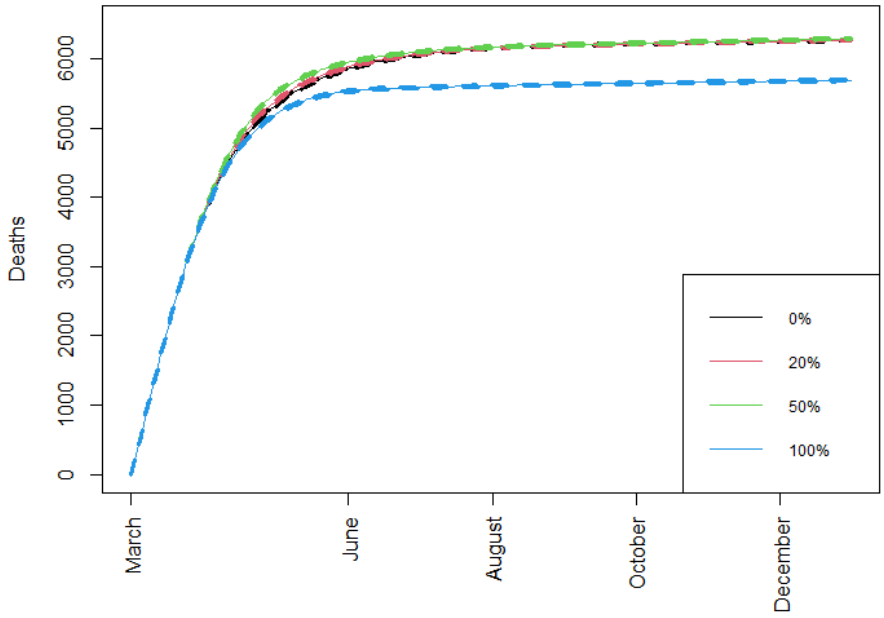


Figure 2.2 Cumulative number of deaths over time when different percentages of doses are allocated under strategy II, with $R_0 = 1.2$, immunity drop between the two vaccine doses is at the Baseline scenario, and with standard vaccine availability.

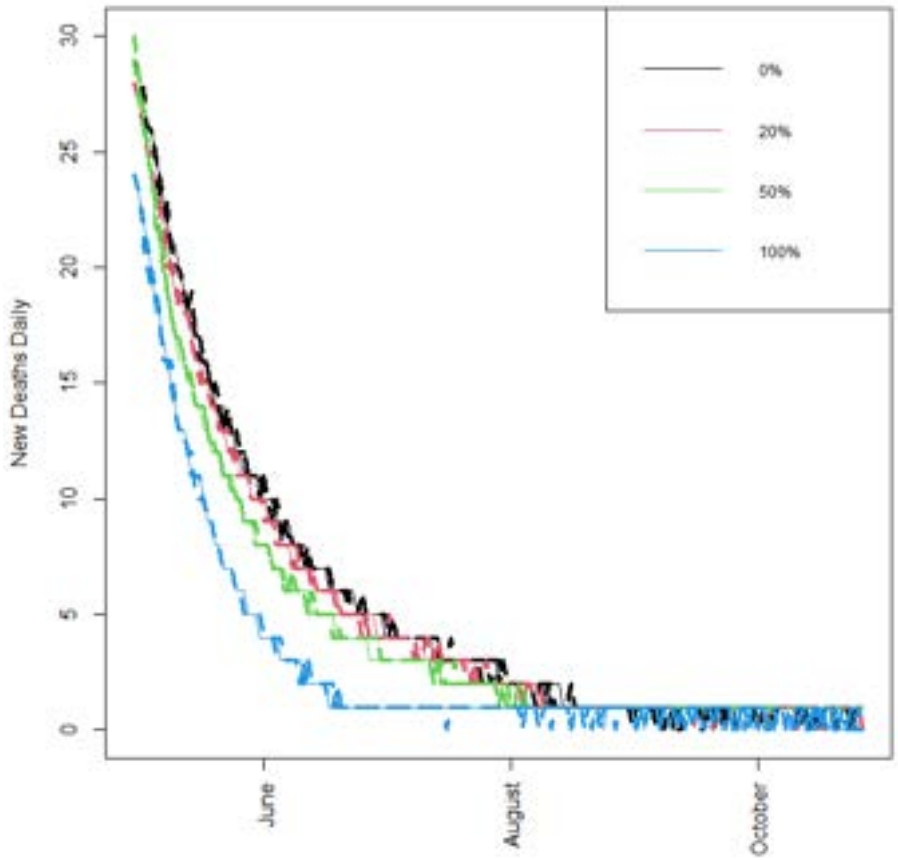


Figure 2.3 Number of new daily deaths, when different percentages of doses are allocated under strategy II, $R_t = 1.2$, immunity drop is at the baseline scenario, and with standard vaccine availability.

2.4 Model assumptions and extra reults

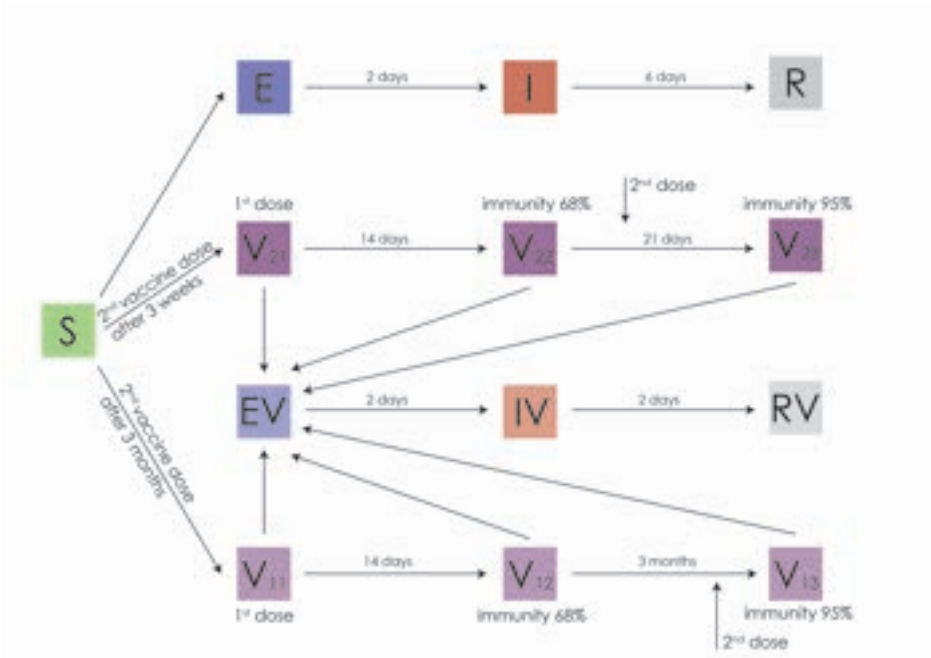


Figure 2.4 Schematic representation of the S(V)EIR epidemic model for the optimistic scenario of immunity waning.

Table 2.2 Age specific infection-fatality-ratios

Age specific infection fatality ratios (IFR)	0-17: 0.00003 18-39: 0.00020 40-64: 0.00500 65+: 0.05400	Data from CDC (2020).
Age specific infection fatality ratios (IFR) for vaccinated people	0-17: 01.5e-06 18-39: 1.0e-05 40-64: 2.5e-04 65+: 2.7e-03	We consider vaccinated people are 95% less probable of dying (Haas et al., 2021).

Table 2.3 Assumptions for the S(V)EIR model

Parameters	Value	Comments-References
R_t	1.2 1.4	Set to assess different levels of transmission. R_t is calculated as the largest eigenvalue of the next generation matrix, using an appropriate contact matrix (see below). We consider reduced infection probability for children by 48% (Koh et al., 2020).
Total population	10816287	Data from the Greek Statistics Authority (ELSTAT, 2021) .
Population by age goup	0-17: 1908003 (17.6%) 18-39: 3200713 (29.5%) 40-64: 3539972 (32.7%) 65+: 2167599 (20%)	Data from ELSTAT (2021). $S_{initial} = [N - R_{initial} - I_{initial}]$ $R_{initial} \sim 620000$ $I_{initial} \sim 60000$.
Medical personnel population	Around 250000	Rough estimate form data from ELSTAT (2021)
Exposed period	2 days for non-vaccinated and vaccinated people	Based on an average incubation time of approximately 5 days (Lauer et al., 2020; Li et al., 2020b) (Liu et al., 2020) and assuming that infectiousness starts approximately 2 days prior to the occurrence of symptoms . (Ganyani et al., 2020; He et al., 2020) (Li et al., 2020a).
Duration of infectious period for non-vaccinated people	6 days	Serial interval of approximately 6 days (Bi et al., 2020; Cereda et al., 2021) (Lavezzo et al., 2020).
Duration of infectious period for vaccinated persons	3 days (worst case scenario) 2 days (baseline scenario and optimistic scenario)	Assuming, that vaccinations decrease the infectious period to one third (baseline scenario and optimistic scenario regarding vaccine efficiency) and to one half (worst case scenario). (de Gier et al., 2021) (Singanayagam et al., 2022)

Table 2.4 Matrix of contacts between age groups

	0-17	*18-39*	*40-64*	*65+*
0-17	16.76	4.34	3.59	0.46
18-39	2.55	6.71	3.47	0.86
40-64	1.88	3.10	5.42	0.98
65+	0.41	1.32	1.67	1.41

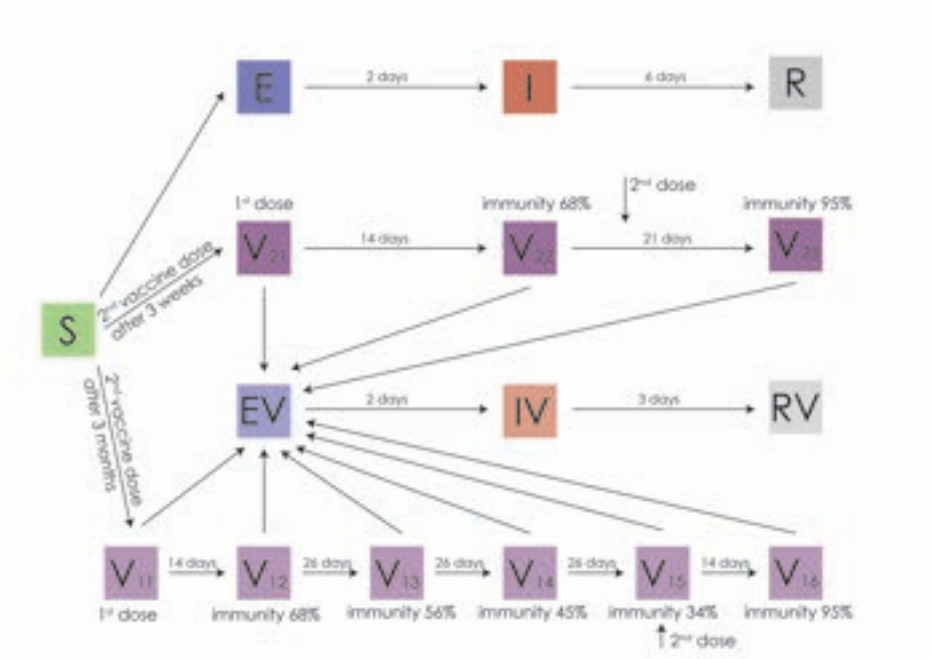


Figure 2.5 Schematic representation of the S(V)EIR epidemic model for the pessimistic scenario of immunity waning.

Table 2.5 Vaccine efficacy assumptions

Parameters related to vaccine efficacy and roll-out		
Vaccine efficacy following the 1st dose and before the 2nd dose	68%	52.4% -92.6% (FDA, 2020; Polack et al., 2020; Voysey et al., 2020) (Baden et al., 2021; Skowronski and De Serres, 2021). [Assuming also reduction in the probability of acquiring infection]. This efficacy is reached 14 days post-vaccination.
Vaccine efficacy after the 2nd dose	95%	Vaccine efficacy in symptomatic infection (FDA, 2020; Polack et al., 2020; Voysey et al., 2020) (Baden et al., 2021; Skowronski and De Serres, 2021). Assuming also reduction in the probability of acquiring infection. This efficacy is reached 14 days post-vaccination.

Table 2.6 Available vaccine doses over time

Period	Doses under normal vaccine availability	Doses under reduced vaccine availability
12/2020	81000	81000
01/2021	350000	35000
02/2021	900000	90000
03/2021	2500000	100000
Q2 2021	5800000	5800000
Q3 2021	6300000	6300000
Q4 2021	3700000	3700000

Table 2.7 Intention to get vaccinated table. Assessed in a sample of 1,097 adults (Sypsa et al., 2021a)

Age group	Total	Probably/Definitely Yes	Probably/Definitely No	Don't know/ Don't answer
18-39	329	193 (58.7 %)	88 (26.8 %)	48 (14.6 %)
40-64	418	288 (68.9 %)	75 (17.9 %)	55 (13.2 %)
65+	350	277 (79.1 %)	36 (10.3 %)	37 (10.6 %)

Table 2.8 Cumulative number of deaths, when 0% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	6 (6-6)	14 (14-14)	15 (15-15)	16 (16-16)	17 (17-17)
18-39	75 (74-75)	143 (143-144)	146 (145-146)	146 (146-146)	146 (146-146)
40-64	1854 (1846-1863)	3649 (3646-3652)	3747 (3745-3749)	3748 (3746-3750)	3749 (3747-3751)
65+	1897 (1886-1906)	2384 (2373-2393)	2425 (2416-2436)	2461 (2451-2471)	2495 (2485-2505)
Total Deaths	3832 (3812-3850)	6190 (6176-6203)	6333 (6321-6346)	6371 (6359-6383)	6407 (6395-6419)
Total Years Lost	73352.5 (72982-73685.5)	134837.5 (134670.5-135044)	138298.5 (138122-138435.5)	138654 (138524-138784)	138995.5 (138865.5-139125.5)

Table 2.9 Cumulative number of deaths, when 20% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	6 (6-6)	14 (14-14)	15 (15-15)	16 (16-16)	17 (17-17)
18-39	75 (74-75)	139 (139-139)	141 (141-141)	141 (141-141)	141 (141-141)
40-64	1852 (1844-1859)	3550 (3547-3552)	3632 (3630-3634)	3634 (3631-3635)	3635 (3632-3636)
65+	1910 (1900-1919)	2498 (2488-2506)	2542 (2533-2551)	2580 (2571-2589)	2617 (2609-2626)
Total Deaths	3843 (3824-3859)	6201 (6188-6211)	6330 (6319-6341)	6371 (6359-6381)	6410 (6399-6420)
Total Years Lost	73383.5 (73020-73656.5)	132451.5 (132291.5-132567.5)	135400 (135277-135523)	135799.5 (135646.5-135892.5)	136162 (136016-136255)

Table 2.10 Cumulative number of deaths, when 50% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	6 (6-6)	13 (13-13)	15 (15-15)	16 (16-16)	17 (17-17)
18-39	74 (74-75)	131 (131-132)	132 (132-132)	132 (132-132)	132 (132-132)
40-64	1847 (1838-1853)	3366 (3363-3369)	3419 (3417-3422)	3422 (3419-3424)	3423 (3420-3425)
65+	1928 (1919-1941)	2723 (2713-2735)	2772 (2762-2784)	2816 (2806-2827)	2858 (2848-2869)
Total Deaths	3855 (3837-3875)	6233 (6220-6249)	6338 (6326-6353)	6386 (6373-6399)	6430 (6417-6443)
Total Years Lost	73306 (72973-73630.5)	128005 (127845-128232.5)	130138.5 (130008.5-130312.5)	130610 (130450-130747)	131007.5 (130847.5-131144.5)

Table 2.11 Cumulative number of deaths, when 100% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	6 (6-6)	13 (13-13)	15 (15-15)	16 (16-16)	17 (17-17)
18-39	75 (74-75)	129 (129-130)	130 (130-130)	130 (130-130)	130 (130-130)
40-64	1854 (1846-1862)	3304 (3300-3309)	3311 (3307-3316)	3313 (3310-3319)	3314 (3311-3319)
65+	1881 (1870-1889)	2278 (2268-2286)	2312 (2302-2320)	2340 (2330-2348)	2367 (2357-2375)
Total Deaths	3816 (3796-3832)	5724 (5710-5738)	5768 (5754-5781)	5799 (5786-5813)	5828 (5815-5841)
Total Years Lost	73240.5 (72870-73536.5)	122923 (122733-123182.5)	123571.5 (123381.5-123777.5)	123901 (123741-124137)	124193.5 (124033.5-124399.5)

Table 2.12 Cumulative number of infections, when 0% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	205692 (201056-210308)	407403 (395412-419507)	449126 (434390-464064)	483022 (465868-500426)	514726 (495223-534530)
18-39	355723 (349647-361850)	657034 (642333-671954)	675324 (658976-691931)	678502 (661464-695861)	681604 (663874-699700)
40-64	354386 (348331-360467)	668880 (654066-683859)	695936 (679082-713033)	699982 (682338-717904)	703266 (684908-721941)
65+	42180 (40320-44094)	57708 (54237-61357)	59220 (55278-63394)	60525 (56142-65195)	61809 (56997-66971)

Table 2.13 Cumulative number of infections, when 20% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	205563 (201024-210191)	403485 (391605-415485)	445046 (430414-459859)	479322 (462237-496643)	511262 (491857-531012)
18-39	355280 (349268-361332)	643554 (628986-658220)	662802 (646528-679252)	668244 (651052-685670)	671562 (653659-689750)
40-64	353940 (347968-360029)	655699 (641042-670620)	682914 (666185-699958)	689238 (671522-707328)	693104 (674617-712012)
65+	42170 (40289-44097)	58765 (55217-62461)	60310 (56284-64541)	61671 (57194-66410)	63013 (58098-68248)

Table 2.14 Cumulative number of infections, when 50% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	205382 (200846-210114)	396362 (384695-408329)	437560 (423185-452361)	472510 (455670-489830)	504882 (485694-524631)
18-39	354592 (348511-360648)	618118 (603982-632366)	638982 (623014-655121)	646856 (629797-664156)	650152 (632378-668213)
40-64	353262 (347216-359440)	630735 (616477-645284)	658194 (641843-674894)	667522 (649974-685486)	671722 (653378-690546)
65+	42128 (40257-44041)	60744 (57067-64557)	62360 (58188-66723)	63826 (59184-68701)	65248 (60157-70642)

Table 2.15 Cumulative number of infections, when 100% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.2.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	205698 (201113-210283)	394808 (383203-406569)	434849 (420571-449360)	470048 (453294-487112)	502502 (483371-521985)
18-39	355640 (349658-361635)	611915 (598340-625688)	638422 (622750-654357)	645310 (628653-662313)	648467 (631115-666220)
40-64	354362 (348317-360332)	623402 (609535-637325)	651674 (635668-667804)	661004 (643853-678331)	664376 (646514-682482)
65+	42111 (40235-44022)	59390 (55683-63234)	61865 (57562-66355)	63053 (58343-68022)	64212 (59091-69658)

Table 2.16 Cumulative number of deaths, when 0% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	12 (12-12)	18 (18-18)	19 (19-19)	20 (20-20)	21 (21-21)
18-39	131 (131-131)	182 (183-182)	182 (183-182)	182 (183-182)	182 (183-182)
40-64	3336 (3336-3337)	4755 (4763-4746)	4778 (4787-4769)	4779 (4788-4770)	4780 (4789-4771)
65+	3925 (3915-3935)	4383 (4377-4390)	4389 (4383-4395)	4395 (4389-4401)	4401 (4395-4406)
Total Deaths	7404 (7394-7415)	9338 (9341-9336)	9368 (9372-9365)	9376 (9380-9373)	9384 (9388-9380)
Total Years Lost	135445.5 (135375.5-135545.5)	184391 (184642.5-184170)	185196.5 (185478-184968.5)	185342 (185623.5-185114)	185487.5 (185769-185252.5)

Table 2.17 Cumulative number of deaths, when 20% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	12 (12-12)	18 (18-18)	19 (19-19)	20 (20-20)	20 (21-20)
18-39	131 (131-131)	178 (179-178)	179 (179-178)	179 (179-178)	179 (179-178)
40-64	3331 (3331-3331)	4665 (4674-4655)	4681 (4690-4671)	4682 (4691-4672)	4683 (4692-4673)
65+	3949 (3937-3962)	4533 (4525-4540)	4541 (4535-4549)	4550 (4543-4556)	4558 (4552-4564)
Total Deaths	7423 (7411-7436)	9394 (9396-9391)	9420 (9423-9417)	9431 (9433-9426)	9440 (9444-9435)
Total Years Lost	135463.5 (135379.5-135554.5)	182527 (182794.5-182276)	183190 (183418-182892.5)	183356.5 (183577.5-183045)	183442.5 (183744-183131)

Table 2.18 Cumulative number of deaths, when 50% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	12 (12-12)	18 (18-18)	19 (19-19)	20 (20-20)	20 (20-20)
18-39	130 (130-130)	172 (172-171)	172 (172-172)	172 (172-172)	172 (172-172)
40-64	3322 (3323-3325)	4502 (4512-4495)	4508 (4518-4501)	4510 (4520-4502)	4511 (4520-4503)
65+	3984 (3970-3994)	4806 (4797-4813)	4821 (4811-4827)	4834 (4825-4840)	4848 (4840-4853)
Total Deaths	7448 (7435-7461)	9498 (9499-9497)	9520 (9520-9519)	9536 (9537-9534)	9551 (9552-9548)
Total Years Lost	135385 (135317-135545)	179227 (179464-179012.5)	179585.5 (179815.5-179417.5)	179810 (180047-179612)	179938 (180152-179733)

Table 2.19 Cumulative number of deaths, when 100% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of deaths					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	12 (12-12)	18 (18-18)	19 (19-19)	20 (20-20)	20 (20-20)
18-39	131 (131-131)	174 (174-173)	174 (174-173)	174 (174-173)	174 (174-173)
40-64	3336 (3336-3336)	4518 (4528-4509)	4524 (4534-4515)	4525 (4535-4517)	4526 (4536-4517)
65+	3895 (3886-3905)	4272 (4266-4279)	4278 (4272-4285)	4280 (4274-4287)	4282 (4276-4289)
Total Deaths	7374 (7365-7384)	8982 (8986-8979)	8995 (8999-8992)	8999 (9003-8997)	9002 (9006-8999)
Total Years Lost	135235.5 (135172.5-135305.5)	176076 (176334-175801.5)	176371.5 (176629.5-176097)	176489 (176747-176244.5)	176533 (176791-176258.5)

Table 2.20 Cumulative number of infections, when 0% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	345538 (339485-351500)	486384 (474256-498460)	515664 (501285-530060)	541945 (525450-558486)	566778 (548240-585388)
18-39	545224 (537607-552807)	742948 (728423-757418)	748351 (732956-763753)	750946 (734926-767032)	753508 (736869-770276)
40-64	561893 (554253-569683)	780841 (765997-796008)	791470 (775350-807974)	794336 (777550-811551)	797174 (779727-815089)
65+	79840 (77276-82458)	93771 (89792-97888)	94662 (90329-99199)	95516 (90822-100454)	96370 (91315-101703)

Table 2.21 Cumulative number of infections, when 20% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	345436 (339414-351463)	483856 (471866-495903)	513237 (498968-527611)	539750 (523376-556309)	564698 (546294-583358)
18-39	544712 (537032-552297)	732614 (718311-746885)	740627 (725203-756066)	744388 (728208-760631)	746981 (730180-763904)
40-64	561348 (553654-569210)	770428 (755638-785499)	782618 (766466-799109)	787262 (770268-804652)	790152 (772497-808256)
65+	79833 (77256-82447)	94773 (90708-98951)	95714 (91280-100312)	96624 (91828-101621)	97528 (92371-102919)

Table 2.22 Cumulative number of infections, when 50% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	345146 (339213-351116)	479122 (467421-490989)	508804 (494847-522989)	535650 (519561-552022)	560812 (542663-579303)
18-39	543699 (536184-551261)	713605 (699927-727448)	726805 (711674-742144)	731324 (715384-747529)	733946 (717374-750836)
40-64	560468 (552777-568128)	751466 (737117-765883)	766841 (750920-782865)	773003 (756131-790043)	775891 (758361-793658)
65+	79766 (77186-82341)	96574 (92398-100835)	97625 (93050-102326)	98616 (93665-103757)	99592 (94274-105169)

Table 2.23 Cumulative number of infections, when 100% of vaccines allocated to ages 18–74, baseline scenario—vaccine availability—R_t = 1.4.

Cumulative number of infections					
Age group	End of March	End of June	End of August	End of October	End of December
0-17	345613 (339630-351653)	480140 (468358-492100)	510030 (495966-524334)	536800 (520574-553305)	561906 (543620-580525)
18-39	545279 (537571-552874)	718430 (704623-732223)	739286 (723640-754978)	742840 (726473-759305)	745458 (728456-762604)
40-64	561980 (554198-569644)	752790 (738433-767005)	776176 (759876-792396)	781252 (764110-798373)	784117 (766319-801958)
65+	79743 (77156-82354)	94982 (90747-99326)	96988 (92242-101914)	97767 (92674-103073)	98526 (93102-104218)

2.5 Discussion

After the vaccination of medical personnel, high-risk individuals, and people aged over 75 years old with a time interval of 3–4 weeks between doses, the strategy of vaccinating the rest of the population with an interval of three months between the two doses (strategy II) can result in a significantly reduced number of deaths and years of life lost. When only 20% or 50% of the available vaccines are distributed using strategy II, the results are not significantly different to strategy I in terms of deaths, although they do provide an improvement in the number of life years saved. In conclusion, rolling out 100% of the available vaccines using the delayed second dose strategy appears to be the most effective option.

In the absence of detailed social contact data between different groups, we accounted for age groups as a surrogate for population composition, and we used the contact rate data between different age-groups from the recent survey (Sypsa et al., 2021b). Therefore, the results reported here offer a conservative assessment since no attempt is made to prioritize individuals with many contacts such as mass transit employees, those working in the hospitality industry, super-markets and so on. Consequently, in practice, the benefits are expected to be even greater if a more targeted approach is adopted.

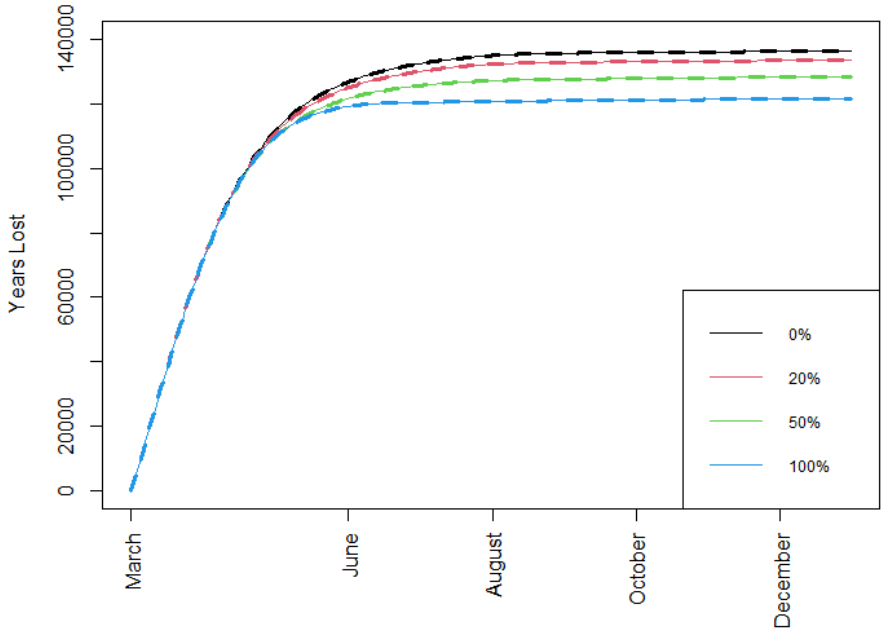


Figure 2.6 Total number of years of life lost when different percentages of doses are allocated under strategy II, $R_t = 1.2$, immunity drop is at the baseline scenario, and with standard vaccine availability.

We used a multitype, age-structured, stochastic epidemic model with constant transmission rate and constant exposed and infectious periods. Although this approach of course has some limitations, they are not expected to materially affect the results. First, in our model, we assumed that vaccine efficacy was mediated by a reduction in infections and not just in clinical disease. Recent modelling studies suggest that, if vaccines reduce symptomatic infection only, then the optimal protection for minimising deaths is obtained by prioritising older individuals (Matrajt et al., 2021b). This assumption is realistic especially in view of recent data suggesting that COVID-19 vaccines are indeed effective in the prevention of infection at least before the occurrence of the Delta strain (Amit et al., 2021; Thompson et al., 2021). Second, we

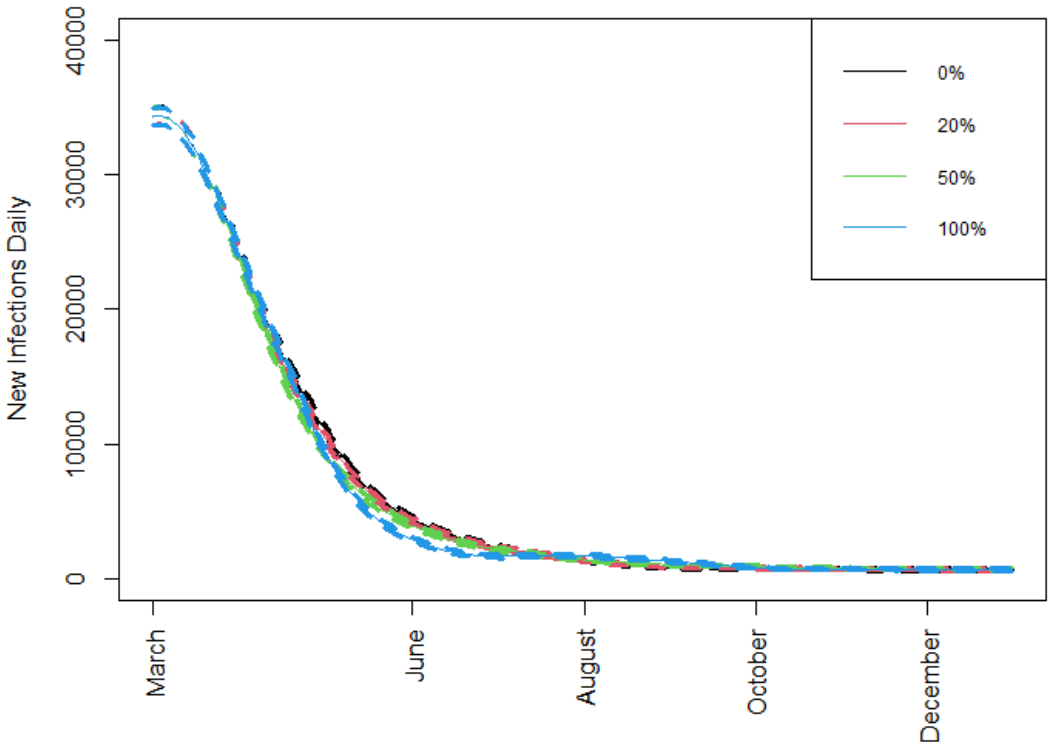


Figure 2.7 New daily infections when different percentages of doses are allocated under strategy II, $R_t = 1.2$, immunity drop is at the baseline scenario, and with standard vaccine availability.

assessed two scenarios for viral transmission rates ($R_t = 1.2$ and $R_t = 1.4$). For higher transmission levels, a recent study similarly found that vaccinating high-risk groups first constituted the optimal use of available vaccines (Matrajt et al., 2021b). On the other hand, moderate transmission levels are a more realistic scenario as most counties continue to implement moderate social distancing measures during vaccination. Alternative scenarios may be considered for the transmission rate, but the overall outcomes are not expected to be substantially influenced as the current assumptions regarding R_t may be thought of as an "average" version of a time-varying rate. In addition, it is known (Andersson and Britton, 2000) that the final size of a stochastic epidemic is invariant to the presence of an exposed period and to different distributional assumptions on the infectious period duration. Hence, these assumptions will not alter the conclusions of this work. Other recent relevant results supporting our assumptions include Tuite et al. (2021); Vasileiou et al. (2021).

The main conclusions of the present study and all relevant assumptions made about vaccine efficacy are in broad agreement with the results obtained using optimization techniques for a model calibrated using data from Italy (Parino et al., 2021), despite the fact that the authors only allowed for the vaccine to protect against transmission and not disease. Recent results from Israel suggest that there is also protection against hospitalization and death

(Amit et al., 2021) and therefore these results may be conservative. Similar conclusions are reported when varying vaccine availability and using alternative efficacy assumptions (Ferreira et al., 2021). Assuming that efficacy remains constant after the first dose, Moghadas et al. (2021) used simulation and showed that the effectiveness of vaccination programmes in reducing infections, hospitalizations and deaths is maximized with a delay of 12 to 15 weeks for both the Pfizer and Moderna vaccines. Similar recommendations on delaying the second dose for individuals below the age of 65 are made in Romero-Brufau et al. (2021).

Although we have chosen to primarily emphasize the results of the proposed approach in terms of quantities of interest in public health, additional gains are to be expected in terms of social and economic aspects of public life by offering faster vaccine coverage to the economically active population. An empirical application of the proposed approach is effectively followed in the United Kingdom and Canada, and the outcome seems to be a significantly faster reduction in SARS-CoV-2 circulation.

Overall, our results clearly indicate that, in the presence of a limited vaccine supply, distributing all available doses with a 3-month intermediate time interval could offer important advantages in terms of public health as well as to the wider society and the economy.

Chapter 3

A stochastic epidemic model for multiphasic infectious diseases

3.1 Introduction

The emergence on early 2020 of Covid-19, an infectious disease caused by the virus SARS-CoV2, has placed health systems around the globe under immense pressure. On March 2020, the World Health Organization declared Covid-19 as a global pandemic, and as of the end of September 2022 more than 6.5 million have died due to illness or complications of it. At the beginning of the pandemic in the absence of available vaccines or suitable medication the majority of governments around the globe resorted to Non-Pharmaceutical-Interventions (NPIs) in an attempt to stop the exponential spreading of the virus and reduce transmissibility. Such NPIs involved

measures like work-from-home policies, school and university closures, stay-at-home guidance for people in high-risk groups and full lockdowns.

These measures had an effect on reducing the transmissibility and resulted in spreading trajectories that could not be properly described by the standard epidemic models due to the resulting multiphasic nature of transmission. The first systematic technique to assess these interventions was due to Flaxman et al. (2020) who proposed a renewal equation model whose infection dynamics were modelled through a multilevel framework incorporating NPIs. We amend this model by inferring the points in time that the transmissibility changes as well as the magnitude of infectiousness in a data-driven manner. We determine the model complexity by using appropriate stochastic processes based upon variations of the Poisson process (PP) and Dirichlet process (DP)-based priors via their stick-breaking constructions (Miller and Harrison (2018); Sethuraman (1994)).

Several models have been proposed in the literature for the estimation of multiphasic infectious diseases, particularly Covid-19. Briefly, a stochastic Susceptible-Exposed-Infectious-Removed (SEIR) model with a regression framework for the effect of the NPIs on transmissibility is used in Knock et al. (2021) while Birrell et al. (2021), Li et al. (2021) and Chatzilena et al. (2022) use stochastic SEIR models where the transmission mechanism is described by a system of non-linear ordinary differential equations and the transmission

rate is modelled by a diffusion process. Modelling the transmission rate as a random walk facilitates gradual and smooth changes in time. A piecewise linear quantile trend model was proposed by Jiang et al. (2021), a kernel-based SIR model distinguishing the different phases of the transmissibility in space was developed by Geng et al. (2021) while Wistuba et al. (2022) incorporated splines to estimate the reproduction number in Germany.

Simpler forms of deterministic and stochastic multiphasic epidemic models have been considered before. In the context of modelling SARS-CoV2 transmission Flaxman et al. (2020) used an approach with fixed number, location and scale of the R_t change. Related work based upon variations of Dirichlet process mixtures is presented in Hu and Geng (2021) and Creswell et al. (2023). In the former, the authors used a Mixture of finite mixtures (MFM) model on a Susceptible-Infected-Recovered-Susceptible model, while in the latter the authors used a suitably modified Pitman-Yor process but only for the scenario of fitting to the observed cases, thus dispensing with the effort to estimate the complete epidemic burden and the suitable adjustment for the reproduction number. The main advantage of the proposed methodology is the intuitive characterization of the epidemic in terms of multiple phases of transmissibility. The number and magnitude of the distinct phases are determined purely by data without explicitly using information about policy changes and NPIs. This approach should be central to a retrospective assess-

ment of the NPIs: an evidence-based method for estimating the timing and effect of those interventions, minimising the risk of introducing several types of bias.

This chapter is organized as follows. In section 3.2 we define the proposed compartmental process, elucidate its equivalence with renewal process-based models and describe the observation regimes of the data. In section 3.3 we complete the model definition by complexity regime. Section 3.4 assesses the proposed models via simulation experiments while section 3.5 contains the application to data from California and New York state, the United Kingdom and Greece. In section 3.7 we present the generalized versions of our proposed models. Information about computation, and software and the convergence of the MCMC can be found in sections 3.8 and 3.9, respectively. The chapter concludes with a discussion.

3.2 Modelling Disease Transmission

3.2.1 Model Definition and Related Characterisations

The methodology for modelling the time-varying disease transmissibility has been implemented under two distinct but equivalent models, the compartmental Susceptible-Infectious-Removed (SIR) model and the seemingly simpler time-since-infection model with population susceptibility reduction. Here we define both models and delineate their equivalence.

In both models we assume that the population has size that equals n and is closed, no births or deaths unrelated to the disease occur during the time frame we observe the epidemic. Also, it is homogeneous in terms of susceptibility, meaning that each individual has the same chance of getting infected and we assume that the population mixes homogeneously, as most of the works on modelling SARS-COV2 transmission. This may be appropriate for large populations such as working at the state or country level since functional central limit theorems can reasonably be applied, e.g. (e.g., Andersson and Britton, 2000) and no data on the household level are typically available. Further extensions of this work can be applied in multitype epidemic models, where the population mixing is separated into different age classes or the transmissibility can be modelled globally and between each household. In the case of Covid19, there is a lack of available data for household transmission and the imposition of NPIs changed the contact patterns of the population.

In the stochastic SIR model, an infected individual makes contact with any other individual on day t at the points of a time-homogeneous Poisson process with time-varying intensity $\frac{\lambda_t}{n}$. This scaling is commonly adopted as it makes the contact rate of meeting any individual independent of the size of the population, as this is a superposition of n independent Poisson processes resulting in rate $\frac{\lambda_t * n}{n} = \lambda_t$ (e.g., Andersson and Britton, 2000). If these (close) contacts of an infected individual occur with a susceptible they

result in an infection. Each individual remains infectious for a random time period Y, which implies that the infectivity of an infected person develops in time without individual variation. An infected individual makes contact with all the susceptibles at day t with rate $S_t * \frac{\lambda_t}{n}$, where S_t is the number of susceptible individuals on day t. This is the rate of the superposition of all the Poisson processes at time points at which an infectious individual makes contact with a susceptible. All Poisson processes in this construction are assumed to be independent. The time-dependent disease reproduction number is defined as $R_t = \lambda_t * E[Y], t = 1, \ldots, T$ where T is the time horizon of the study. The R_t can be described as the expected number of possible infectious contacts an infected individual makes during their infectious period at day t.

For this model the expected number of new infections c_{t+1} at day $t+1$ is given by:

$$E[c_{t+1}] = S_t * \frac{\lambda_t}{n} * I_t * \Delta_{t+1-t}, \tag{3.1}$$

with I_t denoting the active set of infectious individuals:

$$I_t = \sum_{s=0}^{t} \sum_{j=1}^{c_s} P(Y_j > t-s) \tag{3.2}$$

and $P(Y_j > t-s)$ the probability that individual j infected on day s remains infectious on day t. This probability is implicitly determined by the disease

characteristics. Then (3.8) can be rewritten as

$$E[c_{t+1}] = S_t * \frac{R_t}{n} * \frac{\sum_{s=0}^{t}\sum_{j=1}^{c_s} P(Y_j > t-s)}{E[Y]} = \frac{S_t}{n} * R_t * \sum_{s=0}^{t} c_s * g_s(t), \quad (3.3)$$

where $g_s(t) = \frac{P(Y>t-s)}{E[Y]}$ is called the generation interval which defines the time from infection of an individual until the first infection they generate. We assume that all the people infected at day s have the same probability of still being infectious at day t.

Both the probability of a large outbreak and its final size depend on the distribution of the reproduction number and not the generation interval. The generation interval determines how fast the epidemic will grow. The epidemic curve increases at an exponential rate defined by the Malthusian parameter r, which is the solution to the Euler-Lotka equation $R_0 \int_0^\infty exp(-rt) g_s(t) = 1$ (Åke Svensson, 2015). As a result, the generation intervals and the epidemic growth specify the appropriate value of the reproductive number, and therefore, the required control effort to contain the epidemic. This means that health policy need to be informed by infectious disease surveillance systems about these quantities on the required control effort (Wallinga and Lipsitch, 2007). The authors in Åke Svensson (2007) show that sampling schemes and the disease spread dynamics, which are always changing due to the depletion of the susceptible population, are model dependent. The equivalence of the Erlang-distributed SEIR model with the renewal equation models is being

studied in Champredon et al. (2018). Note that equation (3.3) is used in the commonly adopted technique of Cori et al. (2013) for estimating the instantaneous reproduction number. In that approach, the term $\frac{S_t}{n}$ which accounts for the depletion of the susceptible population is ignored since the aim is somewhat different.

One should also consider potential 'superspreading' events when certain individuals infected unusually large numbers of secondary cases, as it was highlighted during the global emergence of severe acute respiratory syndrome (SARS) (Shen et al., 2004; Lipsitch et al., 2003). Having neglected this individual variation and assuming that all the infectors have the same reproduction number, new cases at day t: $c_t \sim Poisson(E[c_t])$. We account for this variability assuming that the individual reproduction number is gamma distributed with mean R_t and dispersion parameter k, yielding $c_t \sim NegativeBinomial(E[c_t], k)$ (Lloyd-Smith et al., 2005). From a modelling perspective, this formulation allows us to model the disease transmission in a more robust way allowing for a higher variance of the data.

The Disease Reproduction Number

The reproduction number R_t is of great practical interest as it is used to assess if the epidemic is growing or shrinking. Here we consider two distinct instances of reproduction number. The effective reproduction number $R_e(t) = S_t * R_t$ describes the expected number of secondary cases generated by an

infectious individual. Then $R_e(t) > 1$ and $R_e(t) < 1$ indicate that the epidemic is growing or shrinking respectively and reducing $R_e(t)$ below unity is the typical target of public health authorities. In contrast, R_t quantifies contacts an infected individual makes during their infectious period at day t that may not always result in new infections, due to mixing with the immune proportion of the population. Therefore, $R_t > 1$ does not necessarily mean that the epidemic is growing due to contacts with a probably immune population, but in a fully susceptible population, the contact rates of the population would be sufficient for an outbreak. The R_t is crucial at the beginning of the epidemic called R_0, when the population is fully susceptible for which results for branching process theory exist and state that the epidemic will become extinct (no more infective individuals that can transmit the disease remaining) when $R_0 \leq 1$.On the other hand if $R_0 > 1$ then the epidemic becomes extinct with probability q where q is the solution to the equation $q = \Pi_\lambda(q)$ and Π_λ is the probability generating function of the contact rate λ and with probability $1-q$ the epidemic results in a major outbreak (Andersson and Britton, 2000). The effective reproduction number $R_e(t)$ is the one that should be targeted by the health authorities to remain below 1 through the implementation of NPIs to reduce the term R_t or vaccination programs to reduce the term S_t. A detailed discussion about reproduction numbers can be found in Pellis et al. (2022).

3.2.2 Observation Regimes

We consider two distinct observation regimes, one where the observed number of cases corresponds to the total number of infections, explained below, and whence the total number of infections is indirectly estimated, outlined in 3.2.2.

Observed Infections

The regime where the total number of infections is observed may be of interest in its own right but may also be used for certain transmissible diseases, for example in the analysis of influenza-like illness data when seroprevalence study information is available. Epidemic models are attractive for analysing such data and are naturally defined in terms of infector-infectee pair and the timing of such events. In reality, however, this type of data is rarely available. Disease monitoring is based on the daily reported infections, which are known to be susceptible to multiple problems, including a time lag between the timing of infection and symptom onset or testing positive.

In the case of Covid-19 a large proportion of the population experiences asymptomatic or mild disease (Ward et al., 2021) leading to severe under-reporting. Inference about the reproduction number can be robust when the reported cases are used if depletion of the susceptible population is accounted for, or if the observed proportion of cases remains constant over time. One

way to validate this assumption is by sequentially performing seroprevalence studies to estimate the true disease prevalence and the proportion of unreported incidences. However, regular information was not available in most countries. In the following subsection, we describe an alternative approach that dispenses with the need for this assumption.

Unobserved Cases

The case where infections may not be directly observed has been studied in a different context by Demiris et al. (2014). In the case of the pandemic, it became immediately apparent that the observed number of infections only partially accounts for the complete epidemic burden. An alternative technique was proposed by Flaxman et al. (2020) where the true cases were estimated by back-calculating infections from the daily reported deaths which are likely less prone to under-reporting. This method has the additional advantage of yielding an estimate of S_t and consequently the total burden of the disease. We adopt this approach for the second level of our model and the daily deaths are linked with the true cases via:

$$\begin{aligned} d_t &\sim NegativeBinomial(E[d_t], k) \\ E[d_t] &= IFR * \sum_{i=0}^{t-1} c_t * \pi(i) \end{aligned} \tag{3.4}$$

Accurate estimates of the infection fatality ratio (IFR) and time-from-infection-to-death distribution ($\pi(i)$) are necessary for estimating incidence, treated here as a latent parameter. The IFR and $\pi(i)$ parameters may be calculated independently from external data or in a single stage, leveraging additional evidence from seroprevalence studies as illustrated in 3.5.

3.3 Epidemic Complexity Determination

The number of phases may be treated as a fixed but unknown integer or as a random quantity to be modelled and estimated from data. We describe two such models in the following two subsections.

3.3.1 Deterministic Number of Phases

One way to perform model selection is to fit an adequate number of models with different complexity and select the best one by using appropriate model information criteria or scoring rules, hoping that the best model is among the ones fitted. In the case of our model with the transmission mechanism described above the term 'model complexity' refers to the number of phases of the transmission of the disease studied. In Flaxman et al. (2020) considered an a-priori selected number of phases. Furthermore, the points in time that the control reproduction number R_t changed were also predefined. The locations of these points were informed by NPIs implemented by each government. The

intuition behind that is that the NPIs have a direct effect on transmissibility. This intuition generates new questions about the time period it takes to observe these changes in the data. Flaxman et al. (2020) also assumed that the reproduction number R_t is a piece-wise constant function. In our proposed models we also consider that R_t is a piece-wise constant function and we amend this transmission mechanism by inferring the location and magnitude of R_t changes, using an appropriate model to perform change-point detection directly from the data. In our methodology, the number of phases K of the epidemic is also predefined. Different fits of the model are examined for different values of K and the model that best describes the data is selected using the Watanabe–Akaike information criterion (WAIC) (Watanabe, 2013) and Leave-one-out cross-validation (LOO) (Vehtari et al., 2017). The model is defined as follows:

$$R_t = \begin{cases} r_1, & t \leq T_1 \\ \ldots & \\ r_{j+1}, & T_j < t \leq T_{j+1} \\ \ldots & \\ r_K, & T_{K-1} < t \leq T \end{cases}$$

$$
\begin{aligned}
r_j &\sim f(\cdot), \quad r_j \in (0,\infty), \quad j = 1,...,K \\
T_{i+1} &= T_i + e_i \\
T_1 &\sim \text{Uniform}\,(3,T) \\
e_i &\sim \text{Uniform}\,(0,100)\,, \quad i = 1,...,K-1
\end{aligned}
\tag{3.5}
$$

3.3.2 Stochastic Number of Phases

Under the Bayesian paradigm, a natural but not trivial way is to treat the model complexity, in our case the number of phases of the epidemic K, as a random parameter, assigning a suitable prior distribution (mostly Poisson or a Categorical distribution, or any distribution in 1,2,...) to it and perform inference through its posterior distribution. The weights of each phase then follow a $Dirichlet(\gamma_1,\gamma_2,..,\gamma_K)$, with $\gamma_i = 1$ not dependent on K, in order for the phases to be a-priori uniform. This approach increases the model complexity but eliminates the need for multiple runs for varying model complexity and the use of information criteria for model selection. This is an area with extended literature with contributions primarily on Bayesian mixture modelling and clustering, as well as in semi-parametric density estimation. Several MCMC methods, such as the 'reversible jump', which was introduced by Green and Richardson (2001), and Richardson and Green (1997), have been used in order to explore the parameter space. The novelty of this general algorithm is that it allows the birth or death of new parameters

at each iteration, but it can be difficult to use since it requires the statistician to design good proposals for the reversible jumps. Although it has been used with great success in many contexts, this difficult characteristic of the algorithm has sustained it from being used in areas beyond clustering.

In our work, we learn the number of phases K via modelling K as a characteristic of particular stochastic processes using generic MCMC algorithms implemented in two widely used software, *Nimble* and *Stan*, facilitating the usage of these epidemic models without the need for reversible jumps. We use Finite Mixture Modelling with an unknown number of parameters by assigning a Poisson process (PP) prior with rate λ. Also as an alternative, we use Bayesian non-parametric methods and specifically the Dirichlet process (DP) prior with scale parameter θ (Ferguson, 1973).

Formal Definition of Poisson process: A Poisson process on a measurable set S is a random countable subset Π of S, such that

- for any disjoint measurable subsets $A_1, A_2, ..., A_n$ of S the random variables

  $N(A_1), N(A_2), ..., N(A_n)$ are independent and

- $N(A)$ has the Poisson distribution with mean rate λ, where $\lambda = \lambda(A)$ lies in $0 \leq \lambda \leq \infty$.

Formal Definition of Dirichlet process: Given a measurable set S, a base probability distribution G and a positive real number θ , the Dirichlet

process $DP(G,\theta)$ is a stochastic process whose realisations drawn from the process are probability distributions over S, such that the following holds. For any measurable finite partition of S, denoted $\{B_i\}_{i=1}^{n}$, if $X \sim DP(G,\theta)$, then $(X(B_1),...,X(B_n)) \sim Dirichlet(\theta G(B_1),...,\theta G(B_n))$.

For both processes, we use the stick-breaking representation. The stick-breaking representation of the PP is presented in Miller and Harrison (2013) and of DP in Sethuraman (1994). We opted to use a Poisson process prior instead of the widely used 'simpler' Poisson prior in Finite Mixture Modelling, in order to account for the time dependence of the epidemic data and allow more phases as time moves forward and new data points appear. This property allows the Finite mixture model to be more consistent with the properties of the DP model.

The Dirichlet process prior has been applied extensively in the bibliography in many Machine Learning applications (Quintana et al., 2020) in order to perform semi-parametric density estimation on data where they tend to repeat in a 'rich get richer' fashion. The distributions realized by a DP prior are almost surely discrete and the scaling parameter θ specifies the level of discretisation. In the limit $\theta \to 0$, the realizations from the DP are all concentrated in a single value, thus allowing us to model even epidemic diseases with only one transmission phase. On the other hand, when $\theta \to \infty$ the realisations become continuous. For the values of θ between the two

limits the realisations are discrete distributions, with a positive probability of ties. We consider these a desirable characteristic in epidemic modelling as it allows for piece-wise constant Reproduction numbers and another way of modelling the seasonality of disease outbreaks.

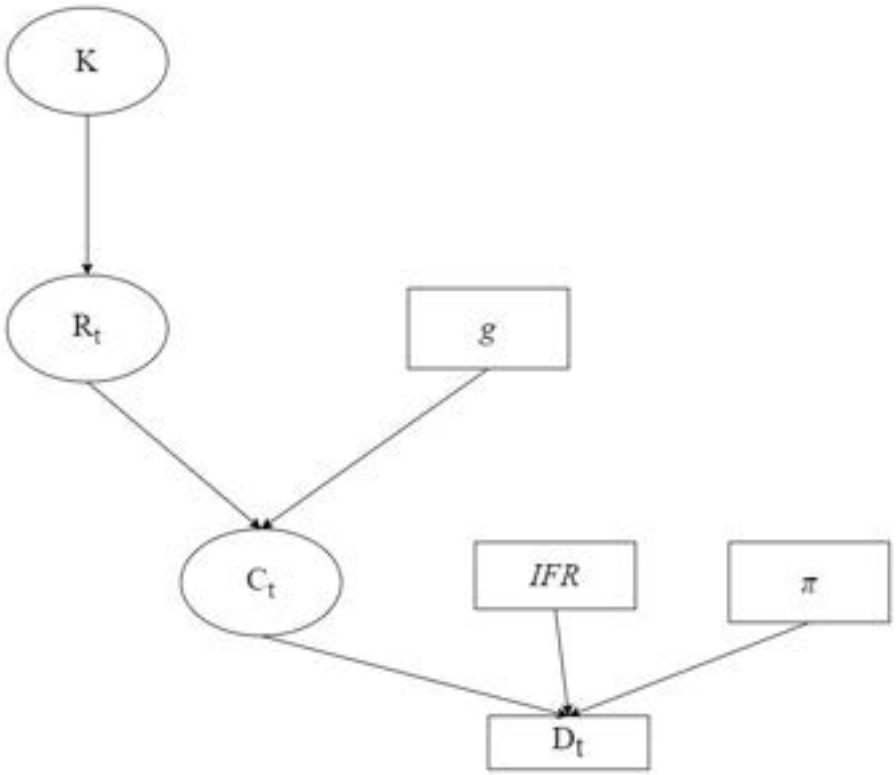


Figure 3.1 Directed acyclic graph of the model. Ellipses denote parameters to be learned by the model. The number of phases K is estimated by the DP/PP model or via model selection criteria.

Estimating the number of phases of the epidemic and the associated location and magnitude of the R_t changes can lead to identifiability problems for R_t and its generative quantities, notably the total number of infections. In order to overcome such issues we explore both a single and a multi-stage

modelling procedure (e.g., Bhatt et al., 2020). In the latter, at the first stage, the latent disease cases are estimated using a Gaussian Process (GP) model and then the medians of these latent cases are treated as data with likelihood given in (3.3). The GP for the estimation of cases is presented in Section 3.6.2.

Poisson Point Process-based Model

We consider that the arrival of new phases in the time horizon (0,T] is driven by a time-homogeneous Poisson process with rate λ, with K growing linearly with time. Hence, following the first epidemic phase, the number, K-1, of new phases follows a Poisson distribution with rate $\lambda * T$ while the duration of each phase a-priori follows an Exponential distribution with rate λ. We follow Miller and Harrison (2013) and use the stick-breaking representation, where the length of the stick is the horizon T we observe the epidemic and each phase breaks length equal to Exponential(λ) from the stick:

$$
\begin{aligned}
R_t &= r_{z_t} \\
r_j &\sim f(\cdot), \quad r_j \in (0,\infty), \quad j = 1,...,K \\
z_t &\sim \text{Categorical}\,(\pi_{1:K}), \quad t = 1,...,T \\
\pi_K &= 1 - \sum_{k=1}^{K} \pi_k, \quad K = min\{j : \sum_{i=1}^{j} T_i \geq T\} \\
\pi_k &= \frac{T_k}{T}, \quad k = 1,...,K-1 \\
T_i &\sim \text{Exponential}\,(\lambda), \quad i = 1,...,K_{max} \\
\lambda &\sim \text{Gamma}\,(0.02, 1)
\end{aligned}
\tag{3.6}
$$

truncating K at $Kmax = 100$, far higher than data-supported estimates.

Dirichlet Process-based Model

An alternative model for the number of phases is based on the DP and its stick-breaking construction:

$$
\begin{aligned}
R_t &= r_{z_t} \\
r_j &\sim f(\cdot), \quad r_j \in (0,\infty), \quad j = 1,...,L \\
z_t &\sim \text{Categorical}\,(w_{1:L}), \quad t = 1,...,T \\
w_L &= \prod_{k<L}(1-v_k), \quad K = \sum_{k=1}^{L} I\{w_k \geq 0\} \\
w_l &= v_l * \prod_{j=1}^{l-1}(1-v_j), \quad l = 2,3,...,L-1 \\
w_1 &= v_1, \quad v_i \sim \text{Beta}\,(1,\theta), \quad i = 1,...,L-1 \\
\theta &\sim \text{Gamma}\,(1,1)
\end{aligned}
\tag{3.7}
$$

where L is the truncation point of the DP, set here to 36. Here K is increasing with the scaling parameter θ.

3.4 Simulation Experiments

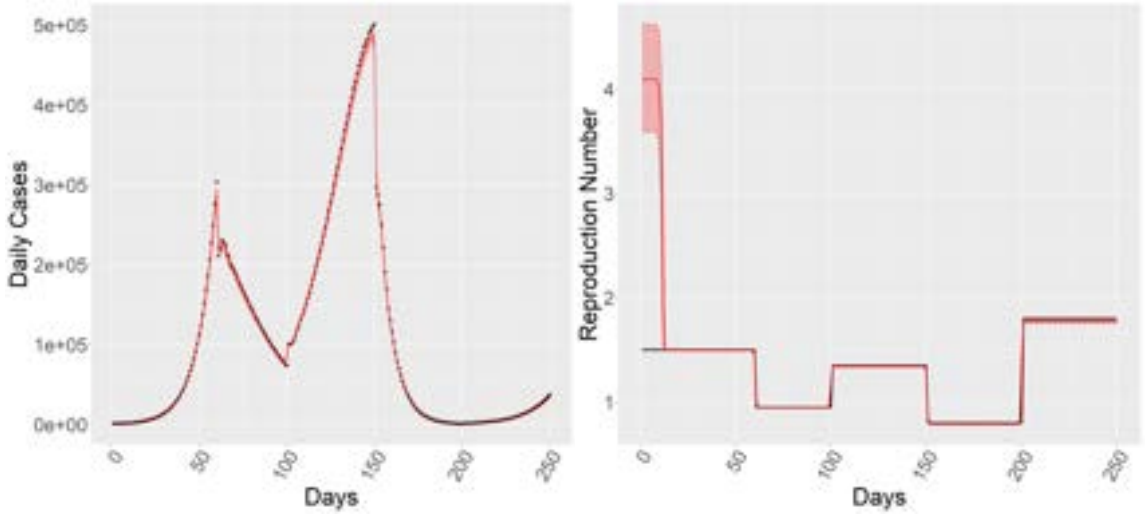
Simultaneously learning the parameters and the dimension of a model is typically a challenging statistical task. We assess the performance of our methods by simulating epidemics of various characteristics for 250 days. The epidemic model defined in (3.2) was used for simulating daily infections and deaths. The population size was set at 10^8 with $IFR = 2\%$. The discretized infectious period and the infection-to-death interval are described in the Section 3.6.1. The epidemic was simulated with 5 distinct increasing/decreasing phases

resembling the observed Covid 19 outbreaks. The time-varying reproduction number was set as follows:

$$R_t = \begin{cases} 1.5, & t \leq 60 \\ 0.95, & 60 < t \leq 100 \\ 1.35, & 100 < t \leq 150 \\ 0.8, & 150 < t \leq 200 \\ 1.8, & 200 < t \leq 250 \end{cases}$$

Using the model in (3.5) and the daily deaths as data the lowest WAIC and LOO selected 5 changepoints. Models with varying (3, 4, 5 and 6) number of changepoints incorrectly identified the first 10 days of the simulation as a distinct phase (Figure 3.2). This can be attributed to the lack of information at the start, a common issue in epidemic models. Following this period the model with 5 changepoints correctly identifies the different epidemic phases, including their timing and magnitude of change. The total daily infections (Figure 3.2) are also accurately recovered. Inference was initiated the day that 10 cumulative deaths were observed. Plots for the other models may be found in Section 3.6.1.

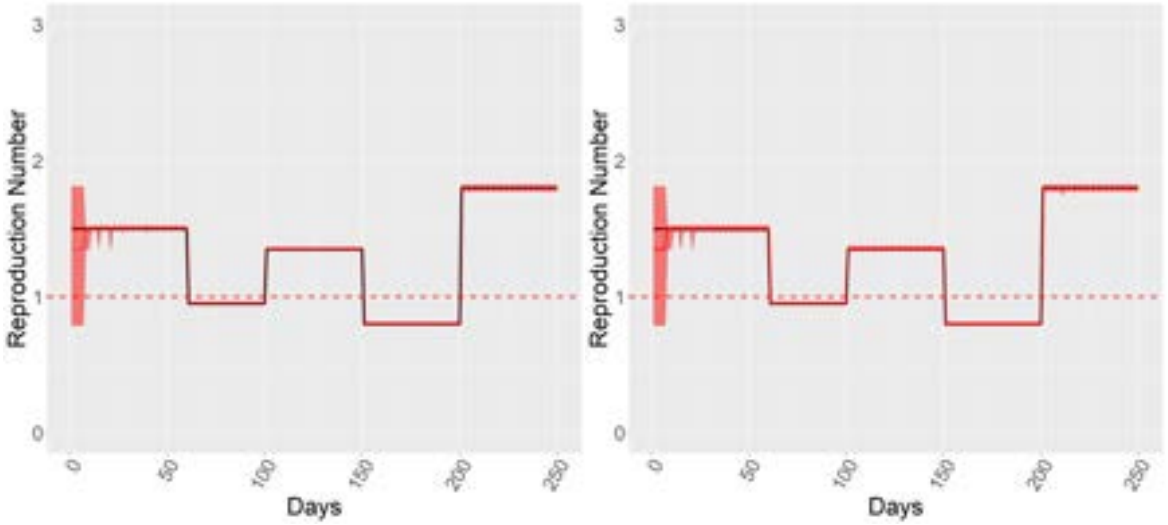
In addition to the findings that the models correctly select the right complexity, it is interesting to summarise the model behaviour when investigating model misspecification. Broadly, these findings may be summarised as follows; when we fix the number of phases to be smaller than the true one then



(a) Simulated (triangles) and estimated daily infections with 95% Cr.I. (line).

(b) Real (solid line) and estimated reproduction number R_t with 95% Cr.I. (dashed line).

Figure 3.2 Simulation and estimates based on observing deaths



(a) Dirichlet process model

(b) Poisson Process model

Figure 3.3 True (solid line) and estimated reproduction number R_t with 95% Cr.I. (dashed line) based on observing infections

the model is correctly recovering the early ones while it is averaging the final phases leading to poorly fitted models. In contrast, when fixing K to be larger than the true one then we essentially recover the true patterns and get a

good fit. Hence, slightly overestimating model complexity is not materially affecting the recovery of the true signal.

When fitting the models with a stochastic number of phases to daily infections, both the PP and DP models are precisely estimating the number of epidemic phases, the time of change and the true R_t value (Figure 3.3). The model was run for 100000 iterations and 8 chains. The analysis based on observing deaths is included in Section 3.6.1. Briefly, the intermediate phases of the epidemic are well estimated while the first and final phases are recovered with noise. The level of smoothing introduced by the cubic spline affects the noisy estimation of the cases; the lower the degrees of freedom the smoother the estimation of cases and subsequently the reproduction number.

3.5 Real-data Application

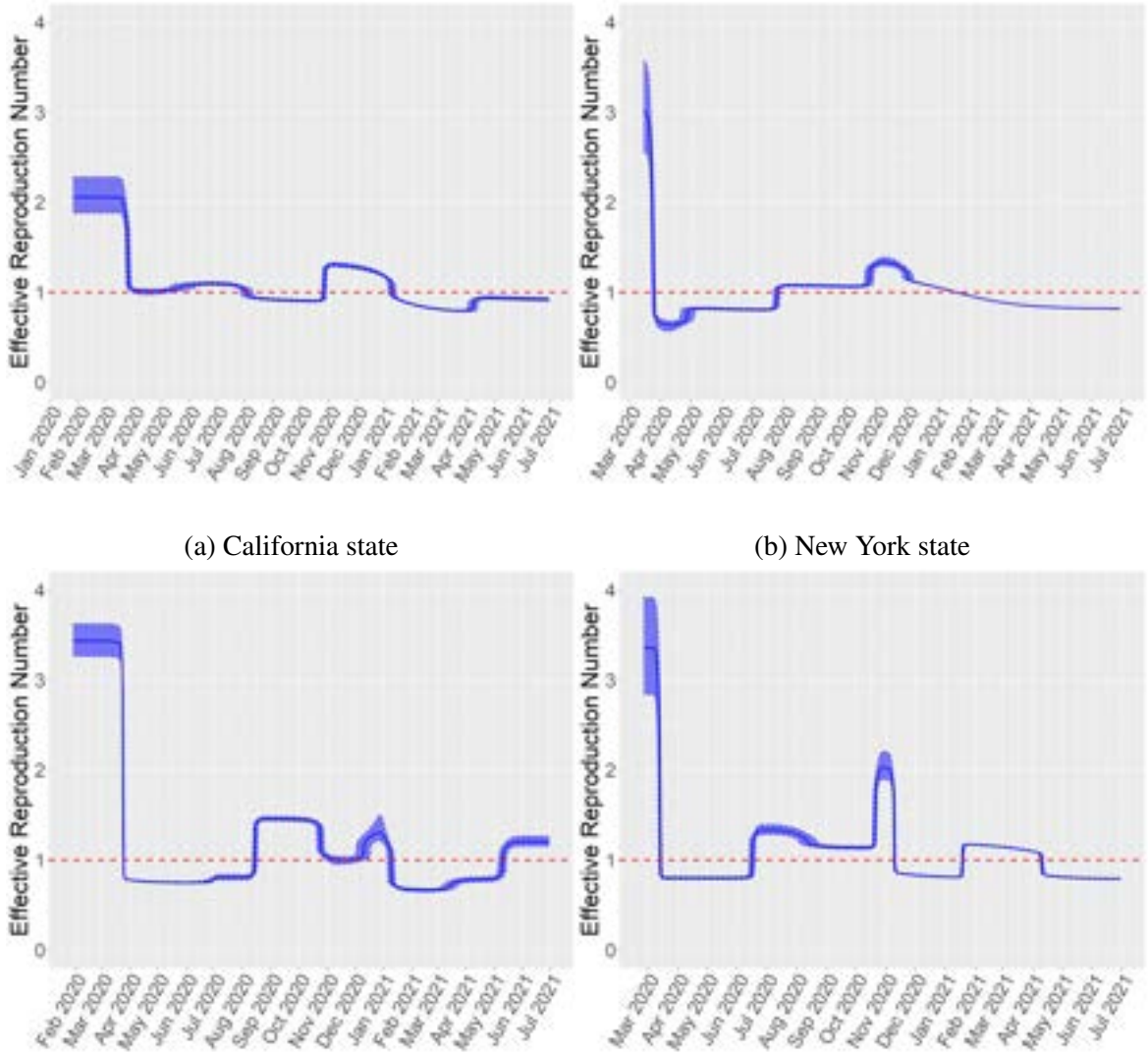
3.5.1 Data Description and Prepocessing

The models were fitted to daily reported deaths from two US states, California and New York and two European countries, the United Kingdom and Greece. The data are accessible from John Hopkins University and ECDC and the time horizon ran to the end of June 2021 when many NPIs were lifted. Due to lack of data availability, the model does not account for reinfections. The age-standardized IFR for each country was informed by the meta-analysis from COVID-19 Forecasting Team (2022) accounting for time, geography

and population characteristics. We allowed the IFR to vary over time, accounting for the age structure of those infected, the burden of health systems and amendments in treating the disease. The infection-to-death time and generation interval were given a Gamma distribution with (mean, standard deviation) set to (19, 8.5) and (6.5, 4.4) days respectively as used by Flaxman et al. (2020).

3.5.2 Analyses and Results

California was one of the first US states reporting cases on the 26th of January, 2020. A state of emergency was declared on March 4, 2020 and mass/social gatherings were banned while a mandatory statewide stay-at-home order was issued on March 19, 2020. We fitted the model to daily deaths and using WAIC/LOO selected 6 changepoints. Figures 3.4 and 3.5 suggest that $R_e(t)$ was reduced after imposing restrictions and fell below the critical value of 1 after April 2020 when school closure was decided for the remainder of the 2019–2020 academic year. The epidemic remained under control until summer 2020 when $R_e(t)$ jumped slightly above 1 following a gradual relaxation of measures. On August 31, 2020, a new set of measures called 'Blueprint for a Safer Economy' was applied and all models show that they were effective, alongside the gained immunity of the population, at reducing the effective reproduction number below one and keeping the epidemic under



(a) California state

(b) New York state

(c) The United Kingdom

(d) Greece

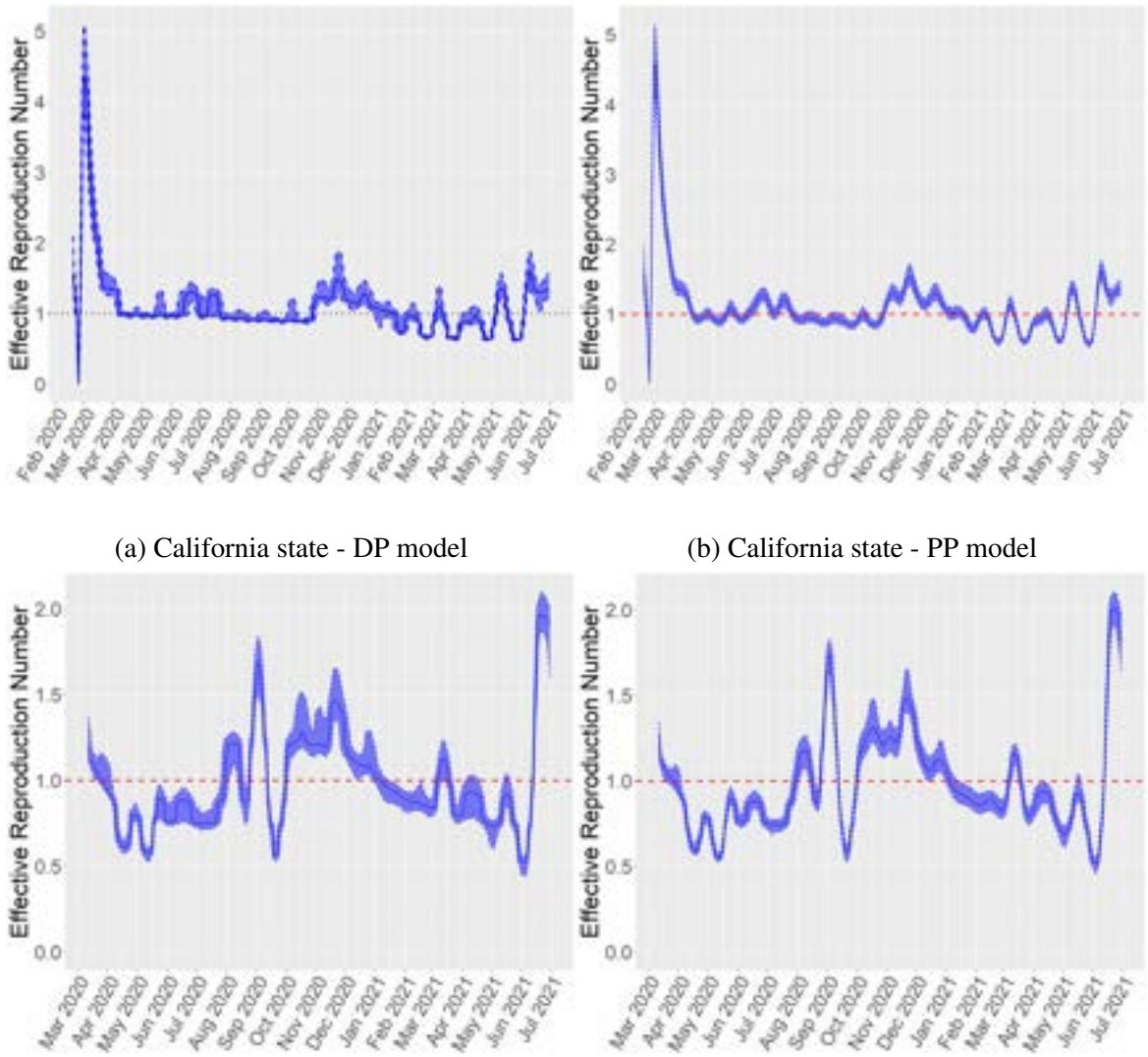
Figure 3.4 Estimation of Effective Reproduction Number $R_e(t)$ with 50% Cr.I. (solid and dashed lines) based on observing deaths, fixed number of phases model.

control until the first half of October 2020. All models estimate a sharp increase in $R_e(t)$, which resulted in an increase in the daily reported cases and deaths between November 2020 and January 2021. Nighttime curfew and regional stay-at-home orders were announced at the start of December 2020 whence $R_e(t)$ remained stable and began declining. The initiation of the

vaccination program on early 2021 brought the epidemic under control with $R_e(t)$ remaining below 1.

New York state had, by April 10 2020, more confirmed cases than any country outside the US and was heavily affected at the start of the pandemic, with daily recorded deaths reaching a thousand in April. On March 15 all New York City schools were closed and on March 20 state-wide stay-at-home order was declared. As a result, the models show a drop of $R_e(t)$ below 1 from mid-March 2020 until August 2020 (Figures 3.4 and 3.5). The best-performing models based on WAIC and LOO had 7 changepoints (8 distinct phases). This model estimates that after the summer of 2020 $R_e(t)$ remained above 1 up until the start of 2021 with a small increase during November and the holiday season. The DP and PP models show similar estimates for $R_e(t)$ (Figure 3.5).

For the *United Kingdom* a model with 8 changepoints was selected by WAIC and LOO. Until early March 2020, when a lockdown was imposed we estimate that $R_t \approx 3.5$ (Figure 3.4). These measures were lifted early June and during the lockdown $R_e(t)$ remained below 1, and therefore under control. After the summer $R_e(t)$ increased above 1 and the so-called rule of six was imposed while on November 5, 2020 the second lockdown was announced. The number of reported deaths was reduced after the initiation of



(a) California state - DP model

(b) California state - PP model

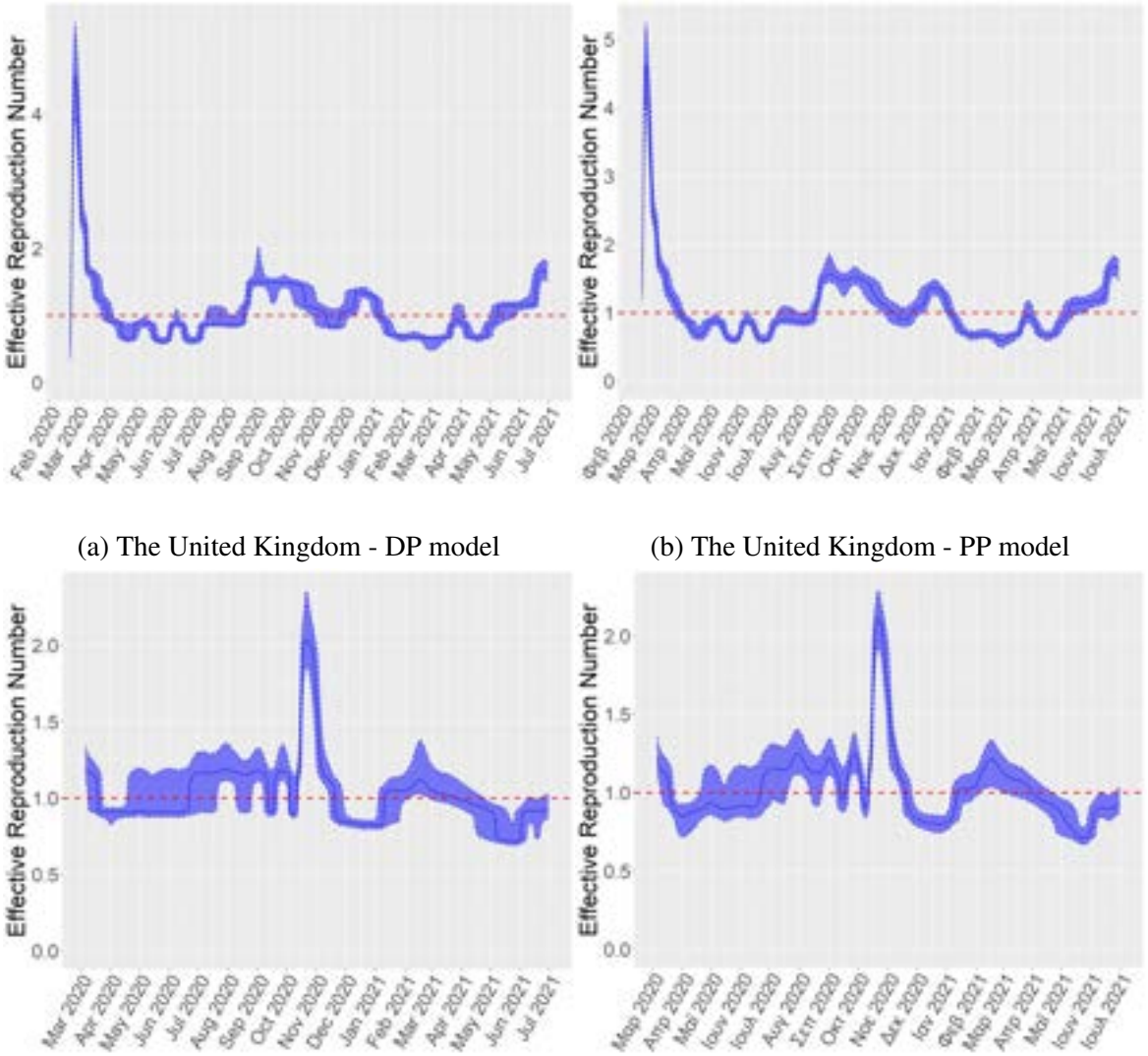
(c) New York state - DP model

(d) New York state - PP

Figure 3.5 Estimation of Effective Reproduction Number $R_e(t)$ with 95% Cr.I. (solid and dashed lines) based on observing deaths, multi-stage approach.

the vaccination program on January 4 2021. Virtually identical estimates for the UK $R_e(t)$ are inferred by the DP and PP models (Figure 3.6).

We conducted an independent (or 'external') validation of the model performance based upon REACT-2, an antibody prevalence study conducted in the UK with the participation of more than 100000 adults (Ward et al.,



(a) The United Kingdom - DP model

(b) The United Kingdom - PP model

(c) Greece - DP model

(d) Greece - PP

Figure 3.6 Estimation of Effective Reproduction Number $R_e(t)$ with 95% Cr.I. (solid and dashed lines) based on observing deaths, multi-stage approach.

2021). This is a unique opportunity as it took place on early July 2020 when waning immunity was unlikely and provides a reasonable estimate of the total disease burden up to that time. The estimated prevalence for the adult population (children were excluded) was 6.0% (95% CI: 5.8, 6.1) and our

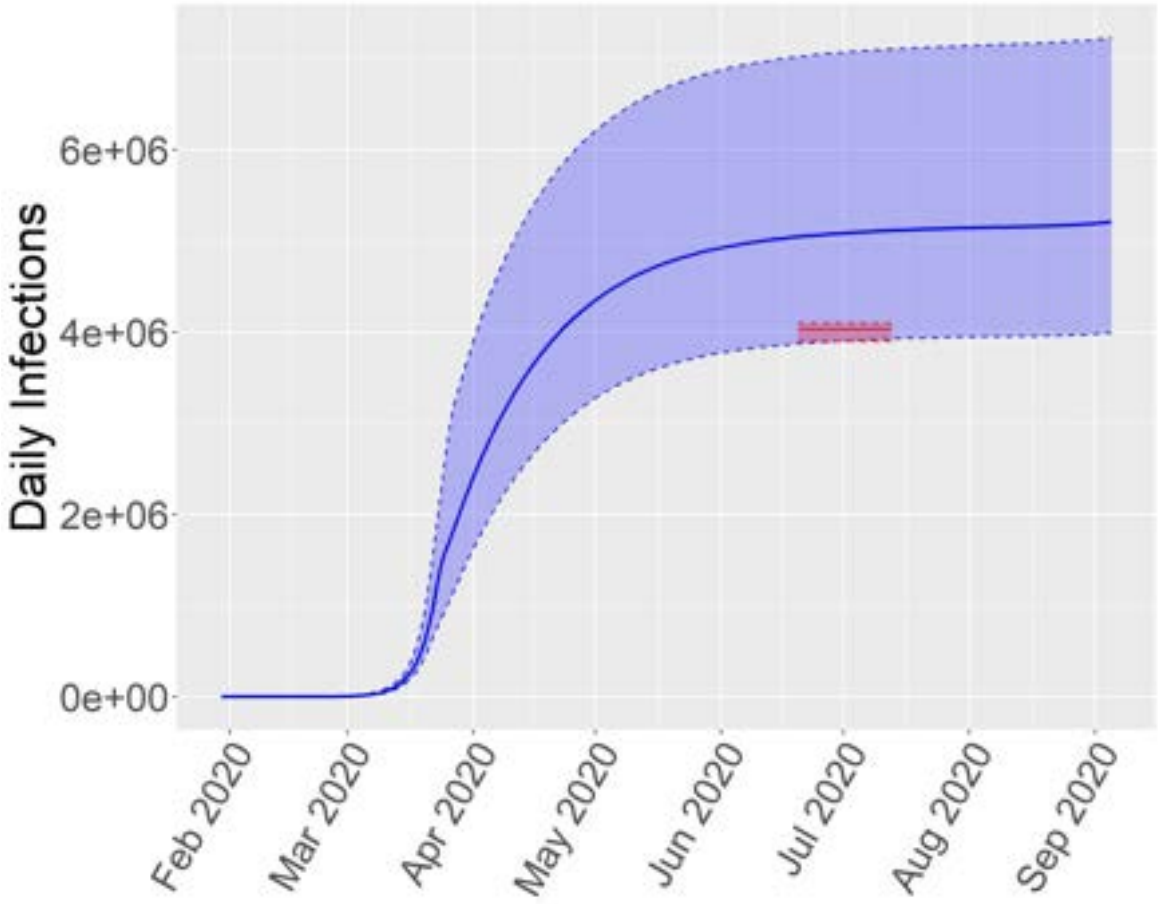
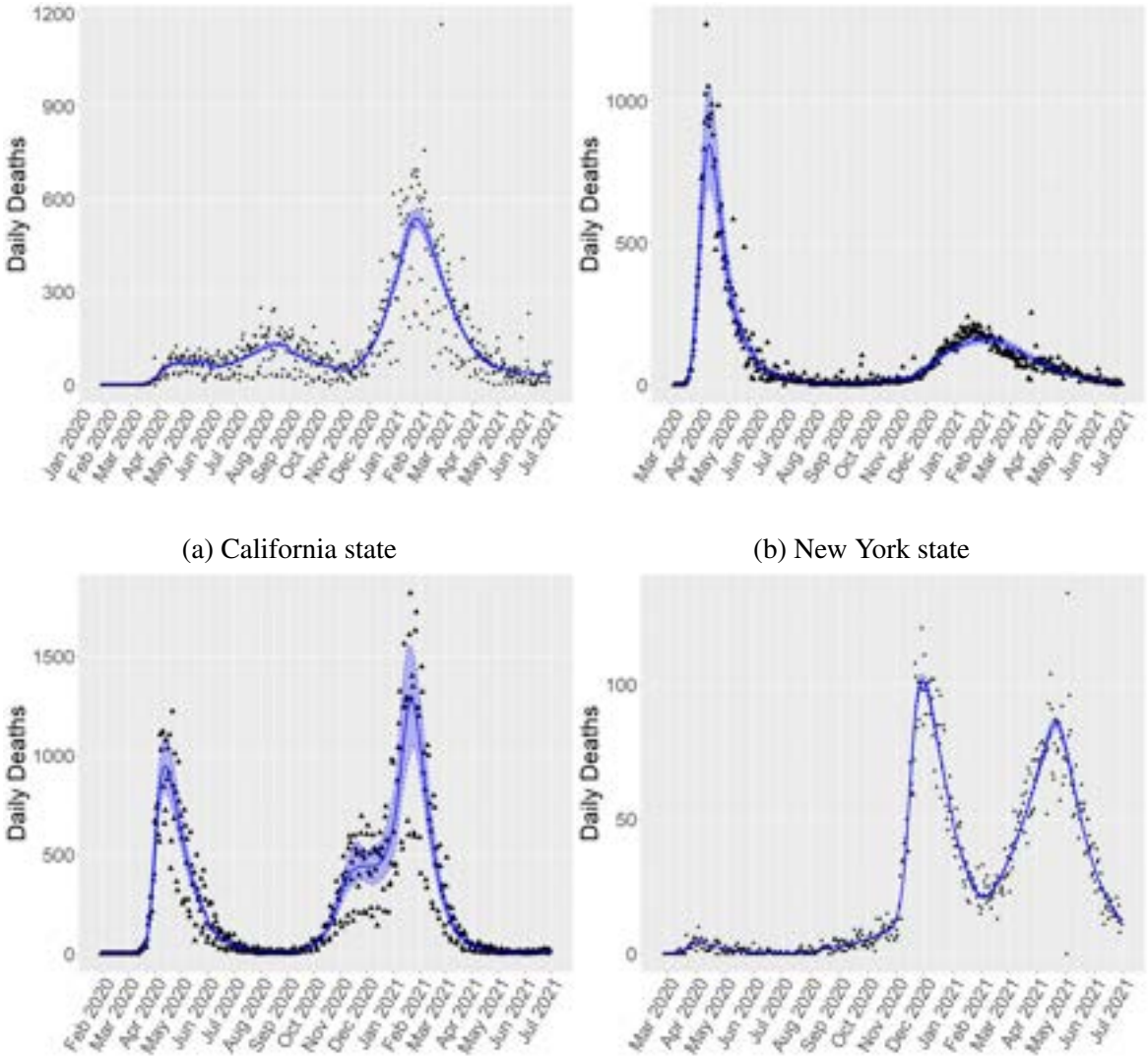


Figure 3.7 Cumulative sum of estimated daily infections with 95% Cr.I. (dashed lines) and the estimation of REACT-2 with 95% C.I. (solid lines) for the United Kingdom

estimate for the whole population is 7.5% (95% Cr.I.: 5.7, 10.) (Figure 3.7) well compatible with that independent estimate.

For *Greece* WAIC and LOO selected the 7-changepoint model. At the starting phase, we estimate $R_e(t) = 3.36$ $(sd = 0.88)$ and a decrease below 1 in the first half of March 2020 (Figure 3.4). On March 10 the government suspended most activities, including educational, shopping and recreational while a week later all nonessential movement was restricted. The $R_e(t)$ estimate remained below 1 until early June 2020 when it increased following the



(a) California state

(b) New York state

(c) The United Kingdom

(d) Greece

Figure 3.8 Reported (triangles) and estimated deaths with 50% Cr.I. (solid and dashed lines) based on observing deaths, fixed number of phases model.

lifting of restrictions. During summer $R_e(t)$ remained over 1 until November 2020 since a case spike in October led to new measures. Similar estimates for the $R_e(t)$ are obtained by the DP and PP models (Figure 3.6).

The results of our simulation experiments corroborate the findings of the application to real data from different areas. The time-ordering of the data

facilitates avoiding label-switching problems typically encountered when fitting mixture models. By selecting the number of phases we capture mortality changes in all the real-world examples (Figure 3.8). The DP and PP models can infer a slightly higher number of phases but the conclusions are not materially affected. This observation is in line with Rousseau and Mengersen (2011) who show a generally stable behaviour of such so-called overfitted mixture models, theoretically verifying the robust behaviour of the developed models. The computation time was similar for the PP and DP models with the DP being faster. More importantly, we get valuable insights on the effectiveness of the measures imposed by the governments. For New York and the UK it appears that the NPIs predate the reductions in transmissibility. California and Greece adopted the measures before a large first wave, like other EU countries and US states. All regions were similar when these measures were relaxed: multiple epidemic waves emerged and the estimated $R_e(t)$ remained above 1.

3.6 Sensitivity analyses

3.6.1 Selecting the number of phases

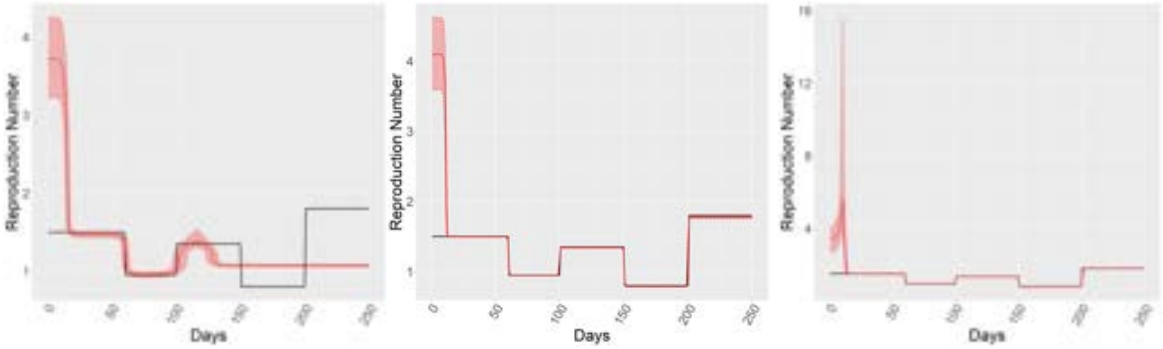
For our simulated datasets we observed that when fixing the number of phases lower than the true one the model is able to identify the first phases of the epidemic but then averages out the transmissibility for the remaining ones,

Fig 3.9. This behaviour to model mis-specification appears reasonable and the averaging of transmissibility has an effect on the daily estimated infections and as a result the models fail to capture properly the second peak of the simulated epidemic (Fig 3.10). Also, the fit of the model on the daily deaths (Fig 3.11) suffers from this type of model mis-specification.

We started training our model for a low number of phases and kept increasing the complexity until the WAIC and LOO values (which initially were decreasing) started rising again. The values of the information criteria are summarized in Table 3.1 for the simulated dataset, and in Tables 3.2, 3.3, 3.4 and 3.5 for California and New York states, and for the United Kingdom and Greece, respectively. Briefly, increasing the model complexity after a certain point does not seem to offer any benefits, since the estimations of the transmissibility (i.e. the effective reproduction number) and of the daily infections remain quite similar (Figures 3.12, 3.13, 3.14 and 3.15), in line with our observations above on the robust behaviour of the model under mis-specification.

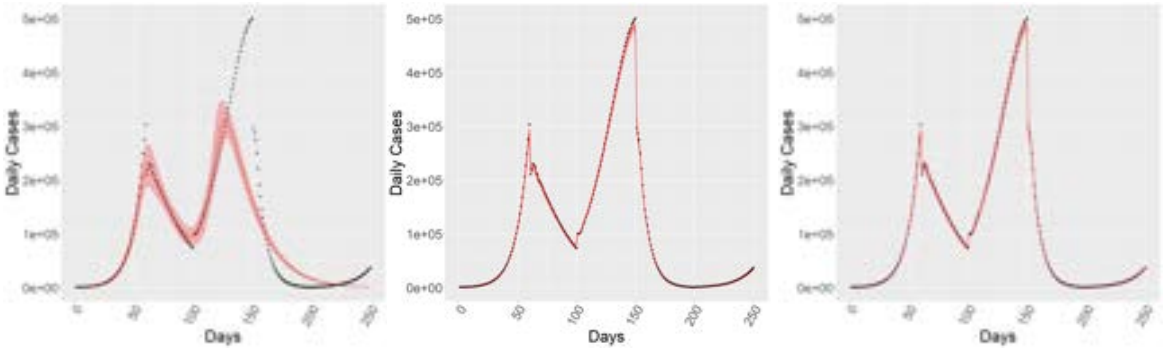
Table 3.1 Fixed number of phases - Simulated dataset

Number of phases	WAIC	LOO
5 phases	3635.0	3635.1
6 phases	2252.9	2253.4
7 phases	2260.6	2261.8



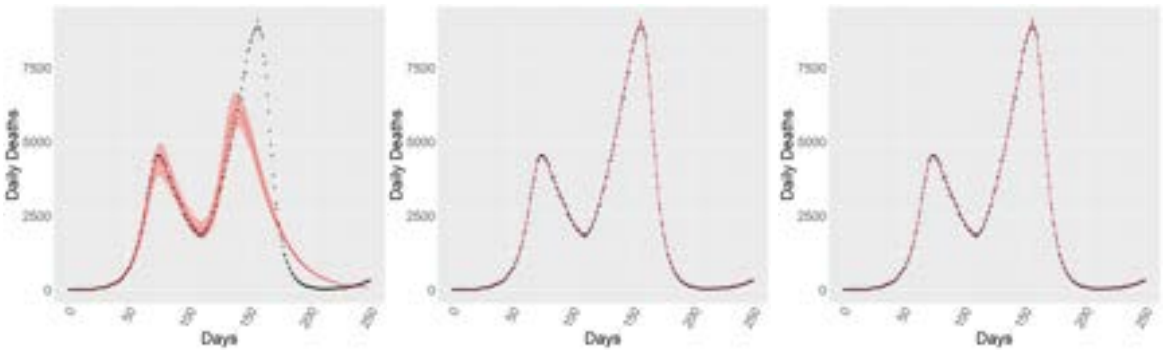
(a) 5 fixed phases (b) 6 fixed phases (c) 7 fixed phases

Figure 3.9 True (black line) and estimated reproduction number R_t with 50% Cr.I. (red line), based on observing deaths in the simulated dataset.



(a) 5 fixed phases (b) 6 fixed phases (c) 7 fixed phases

Figure 3.10 Simulated (black triangles) and estimated daily infections with 50% Cr.I. (red line), based on observing deaths.

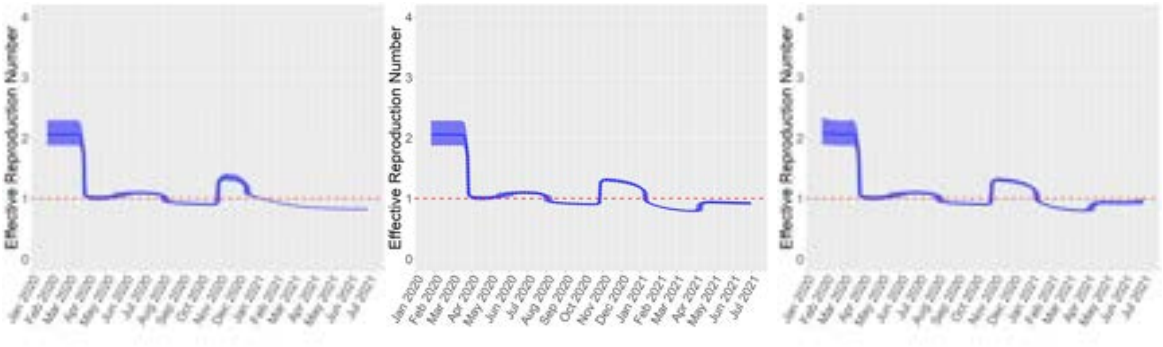


(a) 5 fixed phases (b) 6 fixed phases (c) 7 fixed phases

Figure 3.11 Simulated (black triangles) and estimated daily deaths with 50% Cr.I. (red line).

Table 3.2 Fixed number of phases - California state

Number of phases	WAIC	LOO
6 phases	5099.4	5099.5
7 phases	5064.8	5064.9
8 phases	5066.8	5068.2

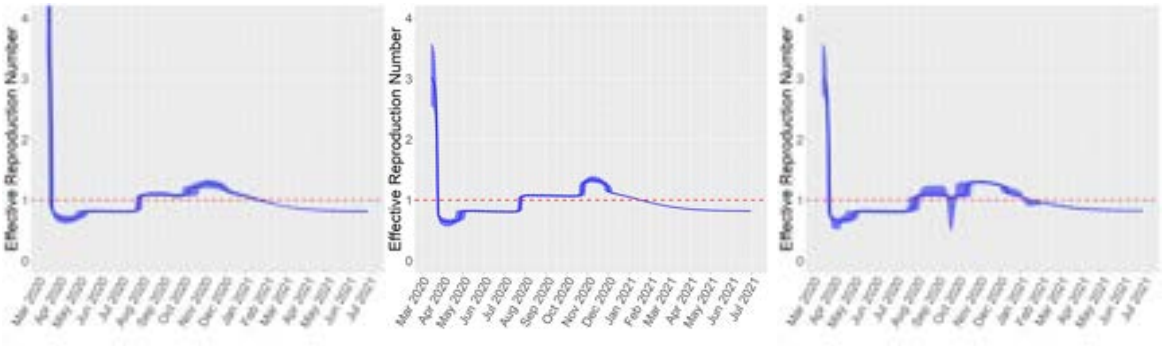


(a) 6 fixed phases (b) 7 fixed phases (c) 8 fixed phases

Figure 3.12 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for the California state.

Table 3.3 Fixed number of phases - New York state

Number of phases	WAIC	LOO
7 phases	4441.5	4439.4
8 phases	4401.6	4401.7
9 phases	4431.0	4415.2

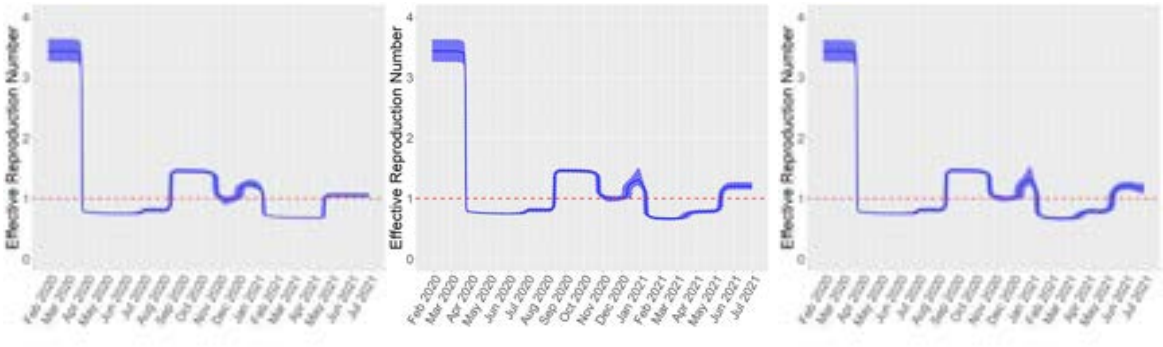


(a) 7 fixed phases (b) 8 fixed phases (c) 9 fixed phases

Figure 3.13 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for the New York state.

Table 3.4 Fixed number of phases - United Kingdom

Number of phases	WAIC	LOO
8 phases	4974.6	4974.7
9 phases	4961.7	4961.8
10 phases	4962.8	4963.0

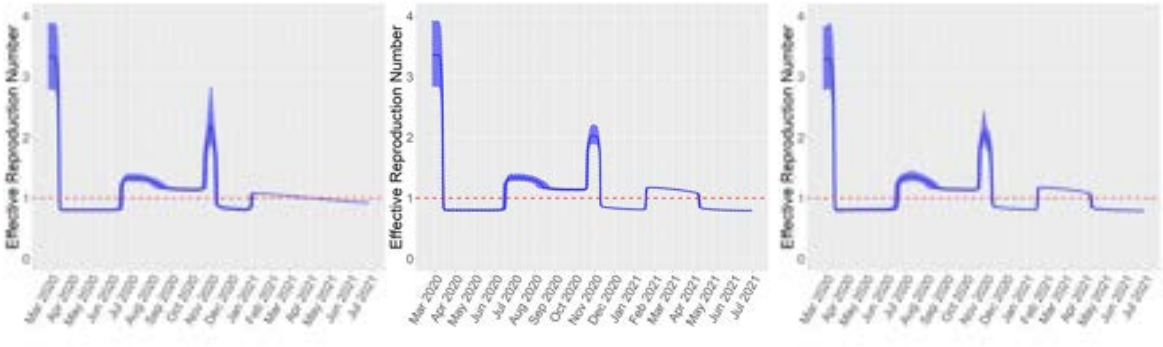


(a) 8 fixed phases (b) 9 fixed phases (c) 10 fixed phases

Figure 3.14 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for the United Kingdom.

Table 3.5 Fixed number of phases - Greece

Number of phases	WAIC	LOO
7 phases	2852.6	2852.7
8 phases	2579.9	2579.9
9 phases	2580.4	2579.8

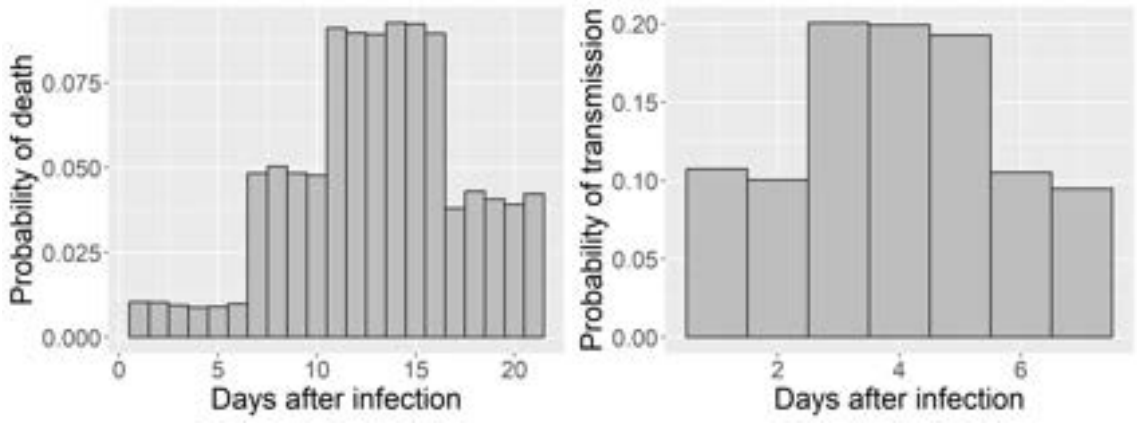


(a) 7 fixed phases (b) 8 fixed phases (c) 9 fixed phases

Figure 3.15 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for Greece.

3.6.2 Selecting the level of smoothing in the multi-stage approach

For the estimation of the transmissibility of the DP and PP models based on daily deaths, in addition to a single-stage procedure we used a multi-stage modelling procedure. In the first stage, the latent disease infections are estimated using a Gaussian Process model and then the medians of these latent infections are treated as data with likelihood given in equation 3.4 and time-from-infection-to-death as shown in Figure 3.16. The model for the estimation of cases is presented below.



(a) Discretized time-from-infection-to-death distribution for simulated experiments.

(b) Discretized generation interval for simulated experiments.

Figure 3.16 Intervals used in the simulation experiments.

$$
\begin{aligned}
d_t &\sim NegativeBinomial(ifr * \sum_{i=0}^{t-1} c_t * \pi(i)) \\
c_t &\sim N(c_{t-1}, s), t = n_{initial} + 1, ..., N \\
c_t &\sim Poisson(\lambda), t = 1, ..., n_{initial} \\
\lambda &\sim \Gamma(1, 0.00001) \\
s &\sim \Gamma(1, 0.00001)
\end{aligned}
\tag{3.8}
$$

The parameter $n_{initial}$ is the total number of initial cases that we seed at the beginning of this Gaussian Process. Due to the seemingly spiky estimates we examined smoothed versions of the estimated cases c_t, by using a cubic smoothing spline fitted to the median estimate of c_t. We applied the multi-stage procedure to the simulated data in order to select an appropriate level of smoothing. The model 3.8 does capture the trend on the daily deaths (Figure 3.17). The estimated cases returned from this model appear in Figure 3.18 (a). The level of smoothing was selected heuristically by visually comparing the estimated cases with the true ones in Figure 3.18, as well as by comparing the estimated R_t with the true one (Figure 3.19). The estimations of the control reproduction numbers from DP and PP models are fairly close for both models. We used 35 degrees of freedom in the cubic splines in both the Dirichlet process and the Poisson process model as lower degrees of freedom resulted in loss of information due to over-smoothing and the elimination of

some epidemic phases while degrees of freedom higher than 45 resulted in noisy estimates.

We fitted model 3.8 to the data of the Covid19 epidemic in California and New York states, the United Kingdom, and Greece up until the first month of the summer of 2021. The fit for each country/state is presented in Figures 3.20, 3.23, 3.26 and 3.29 and the estimated daily infections for various levels of smoothing in Figures 3.21, 3.24, 3.27 and 3.30. As in the simulated dataset, the model does capture daily deaths well and the estimations of the effective reproduction number are similar between the two models (Figures 3.22, 3.25, 3.28 and 3.31).

3.7 Extension of the models

3.7.1 Time inhomogeneous Poisson process

One natural extension of the time homogeneous Poisson process we used to model the number of phases is to assume the rate is not constant for the duration we observe the epidemic. We consider that the new phases (minus 1, $\tilde{K} = K(t) - 1$) arrive at the points of a time-inhomogeneous Poisson process with rate $\lambda(t) = \frac{ve^{vt}}{\mu + e^{vt}}$. Equivalently, the number of phases K - 1 follows a Poisson distribution with rate $\Lambda(t) = \int_0^t \lambda(u)du = log(\mu + e^{vt}) + log(\mu + 1)$. A useful observation from the simulation viewpoint is to notice that for this one dimensional point process the times between the arrival of each

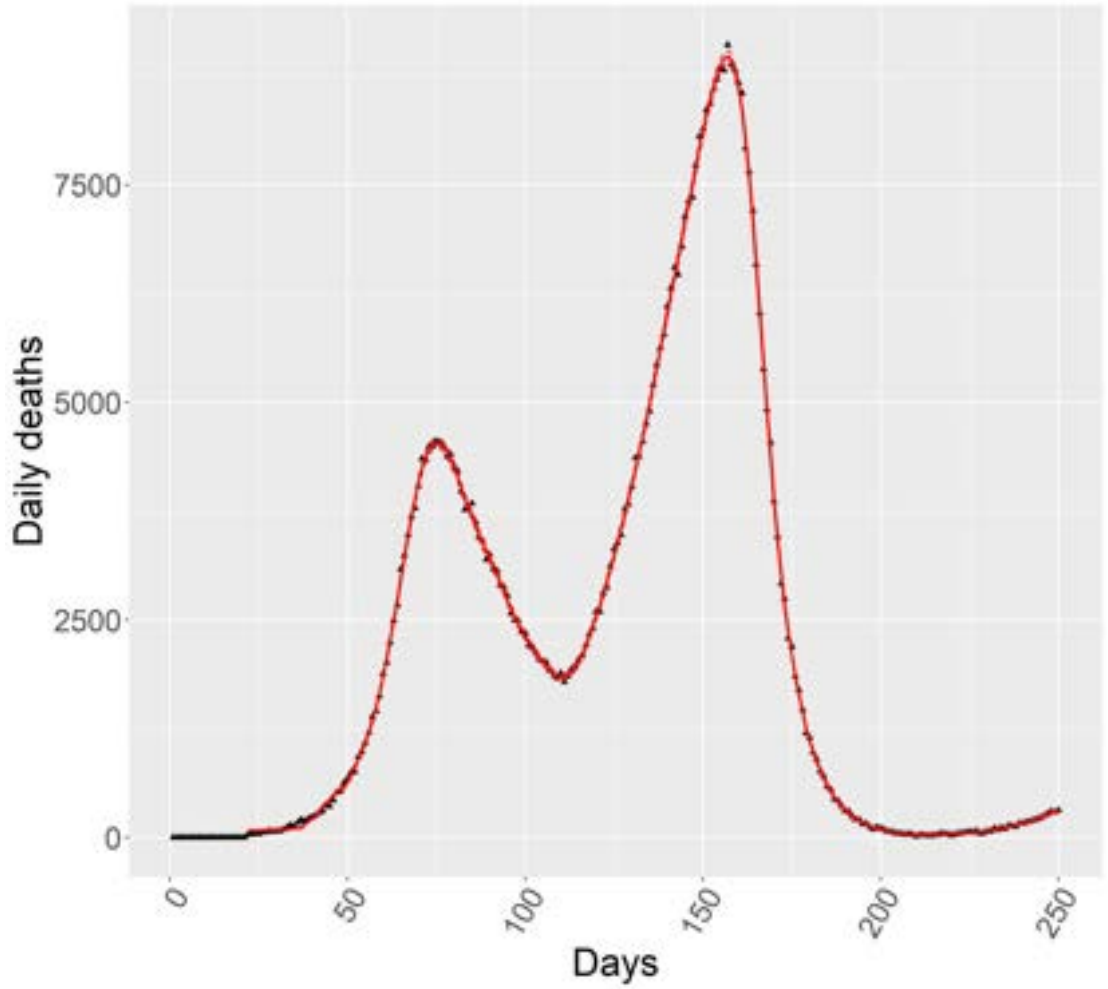
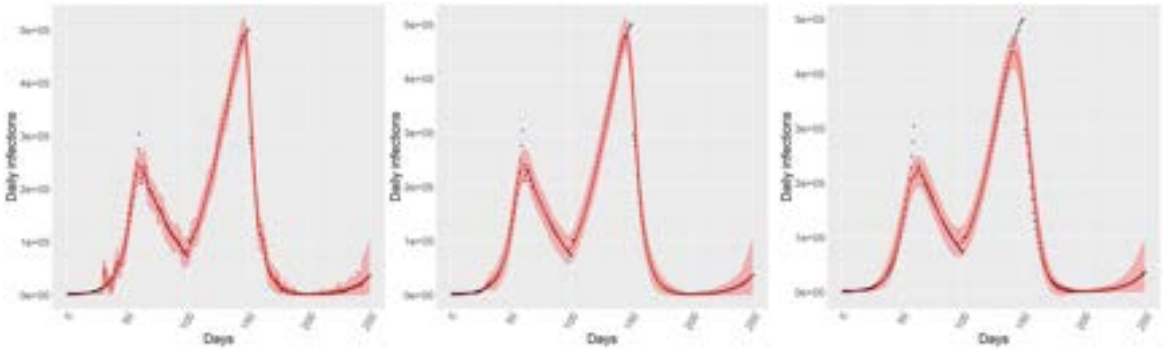


Figure 3.17 Simulated (black triangles) and estimated deaths with 95% Cr.I. (red line) from the model in Equation 3.8.

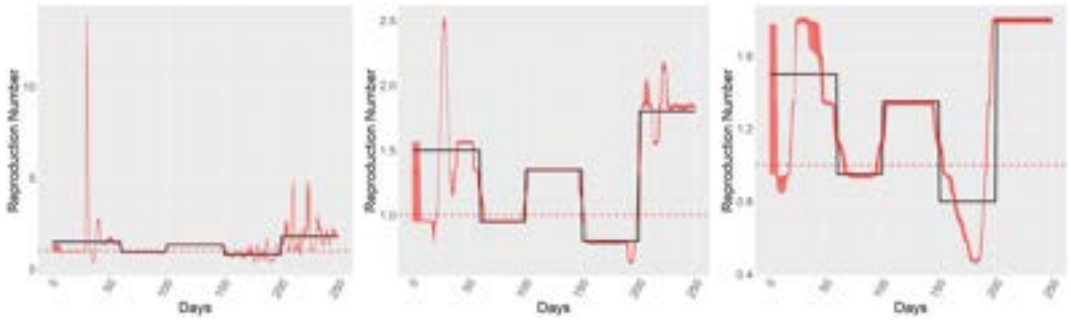


(a) Two stage approach - No smoothing applied

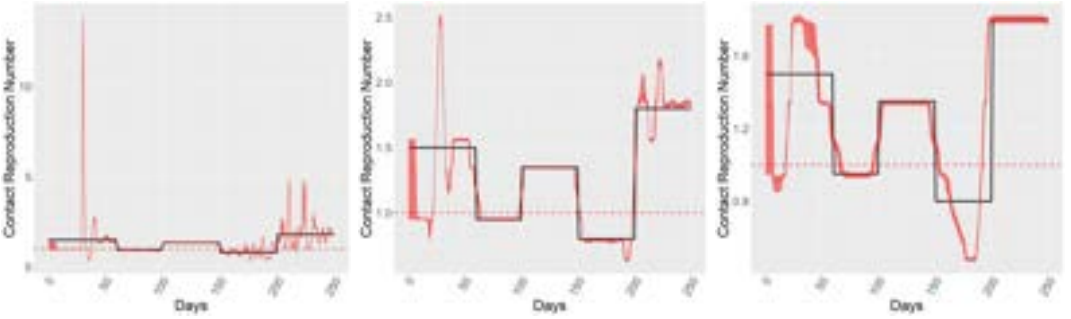
(b) Cubic spline with 35 degrees of freedom

(c) Cubic spline with 15 degrees of freedom

Figure 3.18 Simulated (black triangles) and estimated daily infections with 95% Cr.I. (red line) from the model in Equation 3.8.



(a) DP model - No smoothing applied (b) DP model - Cubic spline with 35 degrees of freedom (c) DP model - Cubic spline with 15 degrees of freedom



(d) PP model - No smoothing applied (e) PP model - Cubic spline with 35 degrees of freedom (f) PP model - Cubic spline with 15 degrees of freedom

Figure 3.19 True (black line) and estimated reproduction number R_t with 50% Cr.I. (red line) based on observing deaths - multi-stage approach.

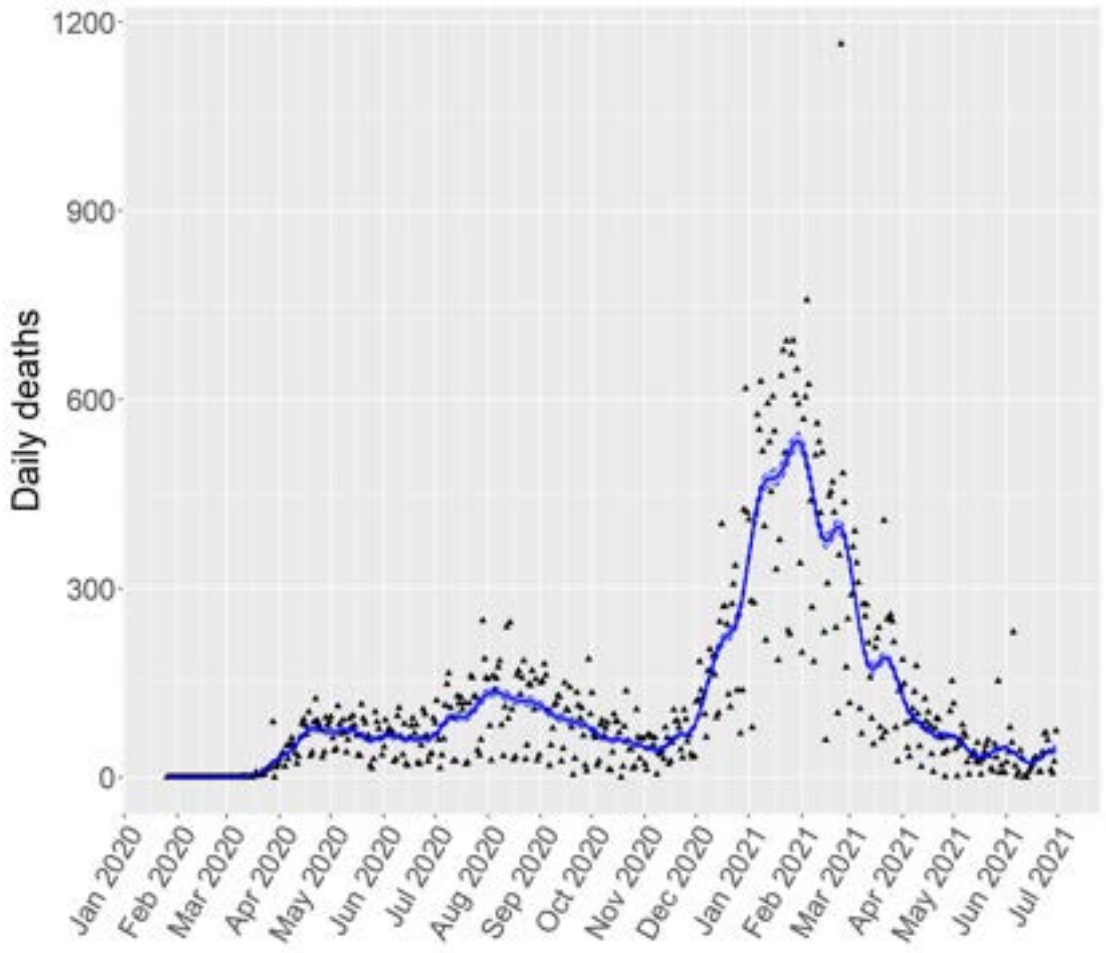
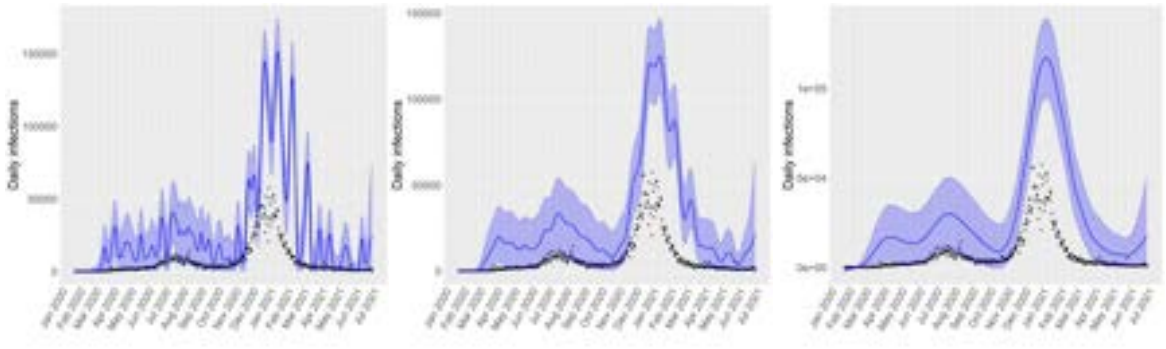
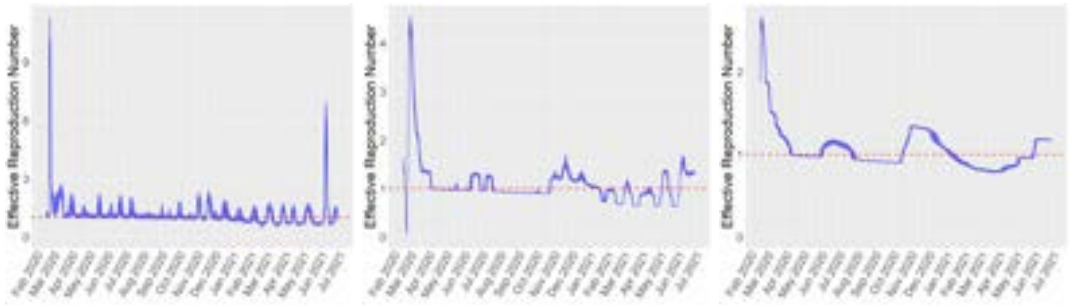


Figure 3.20 Reported (black triangles) and estimated deaths with 95% Cr.I. (blue line) for the California state from the model in Equation 3.8.

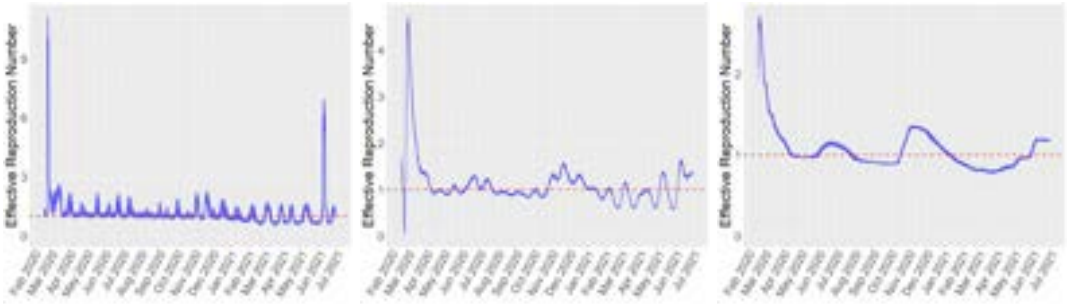


(a) Two stage approach - No smoothing applied (b) Cubic spline with 35 degrees of freedom (c) Cubic spline with 15 degrees of freedom

Figure 3.21 Reported (black triangles) and estimated daily infections with 95% Cr.I. (blue line) for the California state from the model in Equation 3.8.



(a) DP model - No smoothing applied (b) DP model - Cubic spline with 35 degrees of freedom (c) DP model - Cubic spline with 15 degrees of freedom



(d) PP model - No smoothing applied (e) PP model - Cubic spline with 35 degrees of freedom (f) PP model - Cubic spline with 15 degrees of freedom

Figure 3.22 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for the California state based on observing deaths - multi-stage approach.

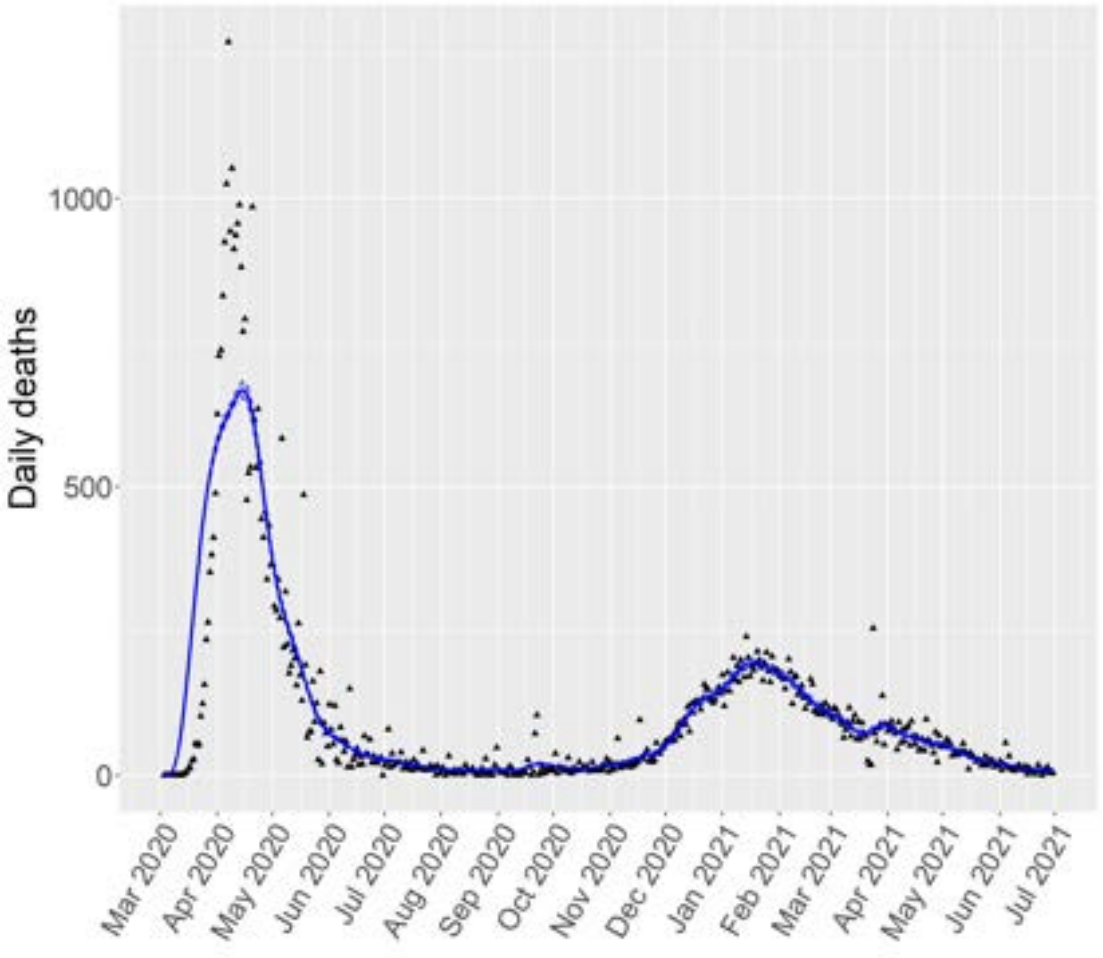
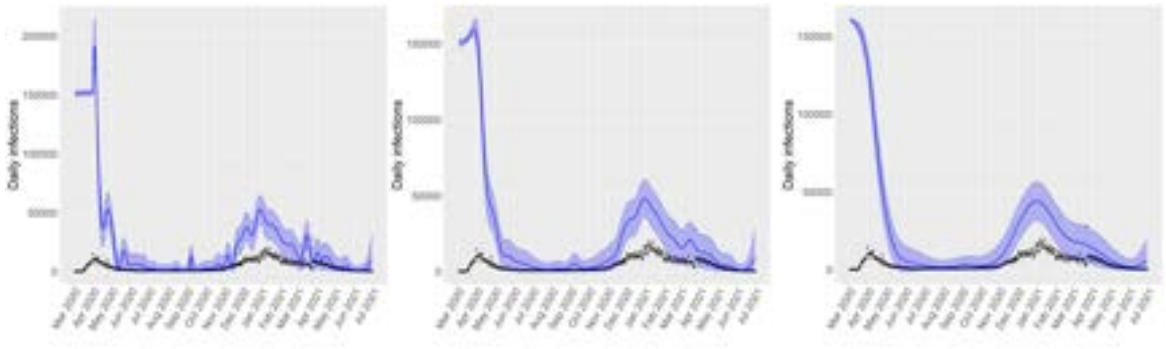


Figure 3.23 Reported (black triangles) and estimated deaths with 95% Cr.I. (blue line) for the New York state from the model in Equation 3.8.

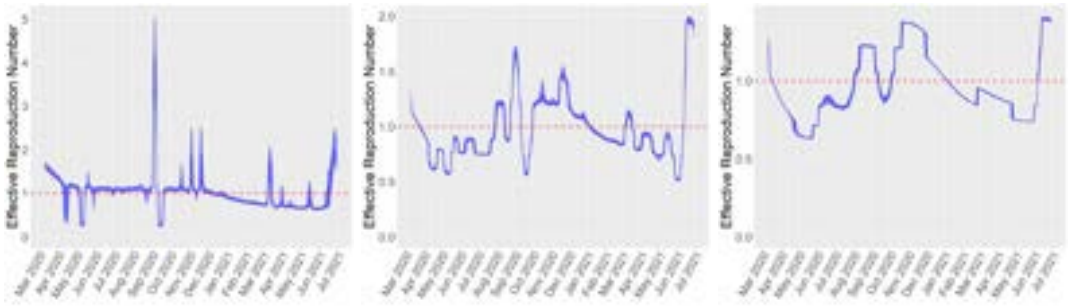


(a) Two stage approach - No smoothing applied

(b) Cubic spline with 35 degrees of freedom

(c) Cubic spline with 15 degrees of freedom

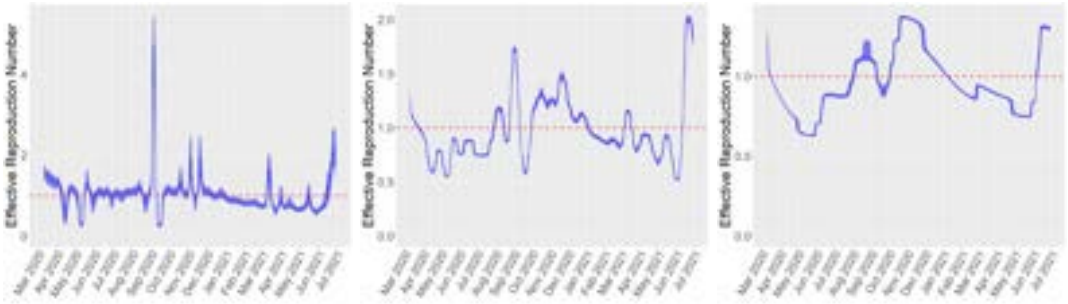
Figure 3.24 Reported (black triangles) and estimated daily infections with 95% Cr.I. (blue line) for the New York state from the model in Equation 3.8.



(a) DP model - No smoothing applied

(b) DP model - Cubic spline with 35 degrees of freedom

(c) DP model - Cubic spline with 15 degrees of freedom



(d) PP model - No smoothing applied

(e) PP model - Cubic spline with 35 degrees of freedom

(f) PP model - Cubic spline with 15 degrees of freedom

Figure 3.25 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for the New York state based on observing deaths - multi-stage approach.

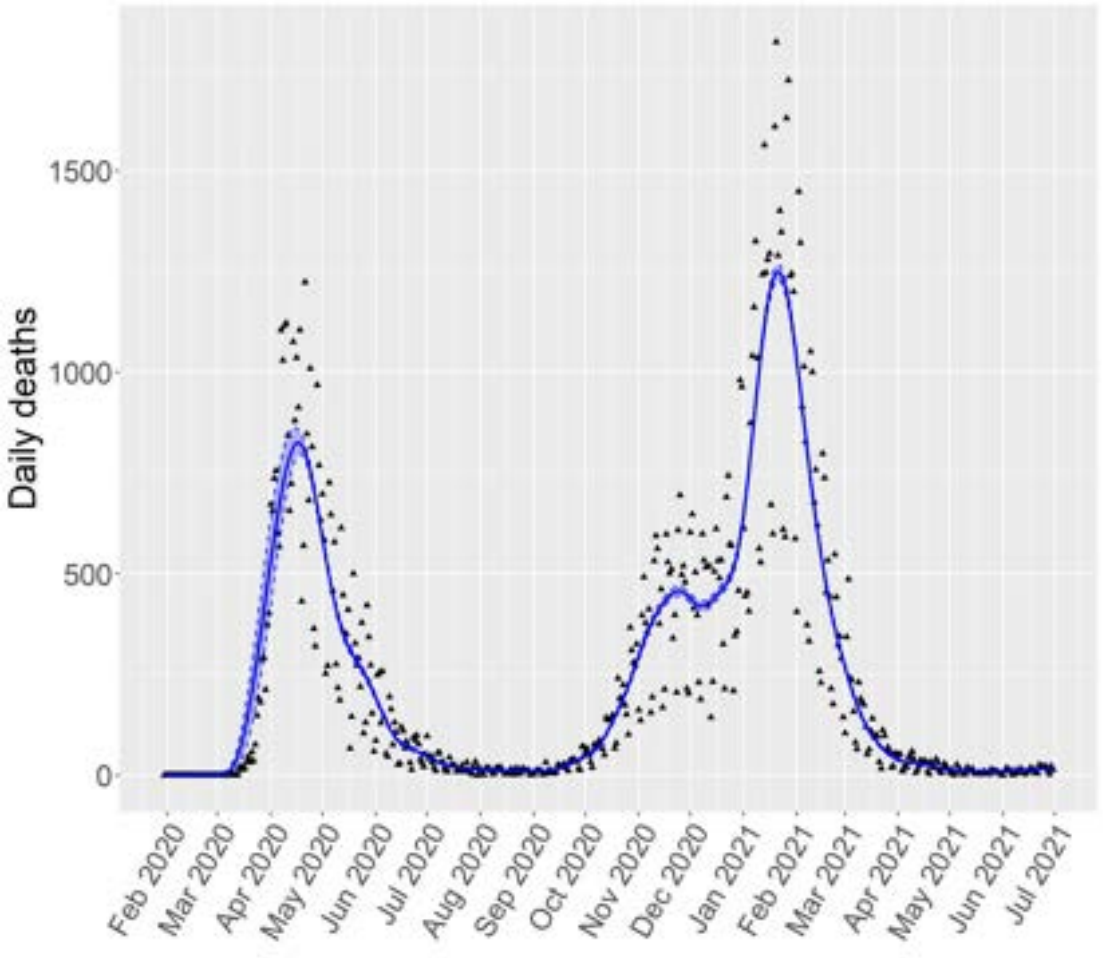
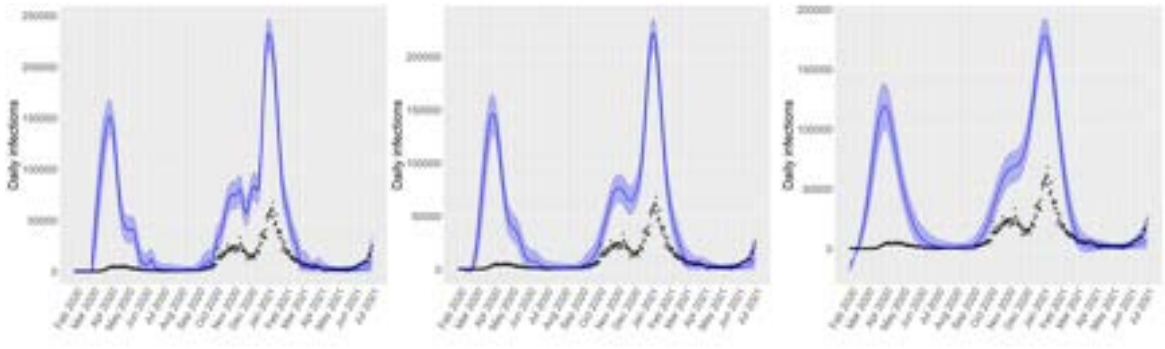


Figure 3.26 Reported (black triangles) and estimated deaths with 95% Cr.I. (blue line) for the United Kingdom from the model in Equation 3.8.



(a) Two stage approach - No smoothing applied (b) Cubic spline with 35 degrees of freedom (c) Cubic spline with 15 degrees of freedom

Figure 3.27 Reported (black triangles) and estimated daily infections with 95% Cr.I. (blue line) for the United Kingdom from the model in Equation 3.8.

(a) DP model - No smoothing applied (b) DP model - Cubic spline with 35 degrees of freedom (c) DP model - Cubic spline with 15 degrees of freedom

(d) PP model - No smoothing applied (e) PP model - Cubic spline with 35 degrees of freedom (f) PP model - Cubic spline with 15 degrees of freedom

Figure 3.28 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for the United Kingdom based on observing deaths - multi-stage approach.

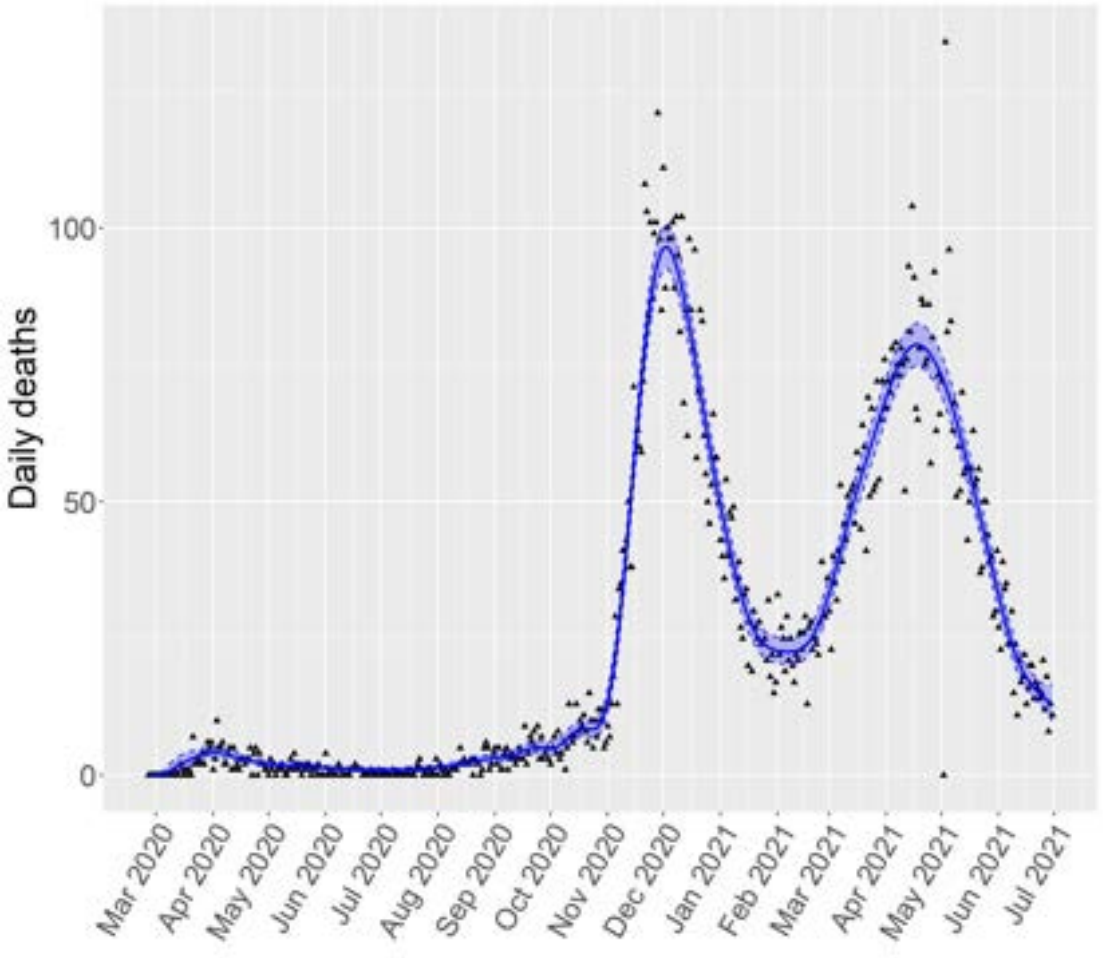
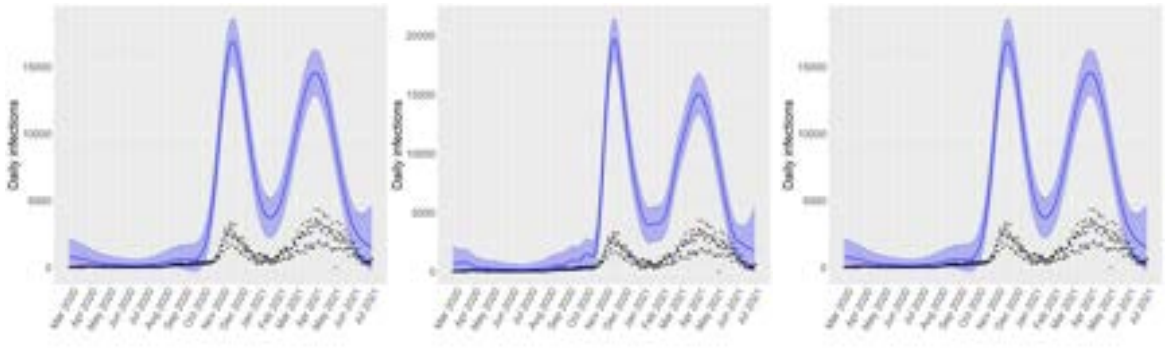
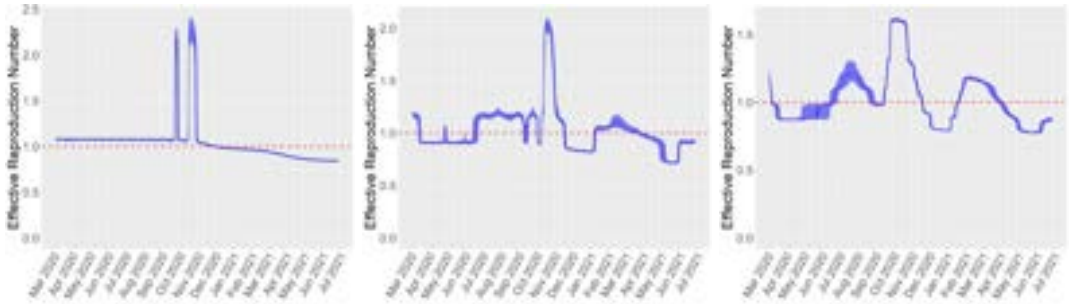


Figure 3.29 Reported (black triangles) and estimated deaths with 95% Cr.I. (blue line) for Greece from the model in Equation 3.8.

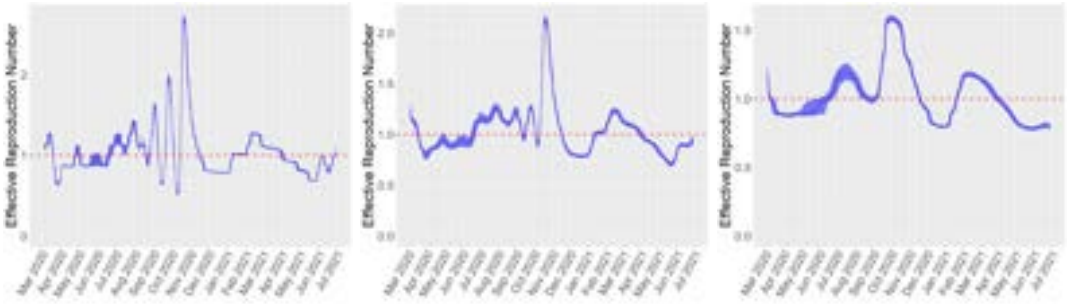


(a) Two stage approach - No smoothing applied (b) Cubic spline with 35 degrees of freedom (c) Cubic spline with 15 degrees of freedom

Figure 3.30 Reported (black triangles) and estimated daily infections with 95% Cr.I. (blue line) for Greece from the model in Equation 3.8.



(a) DP model - No smoothing applied (b) DP model - Cubic spline with 35 degrees of freedom (c) DP model - Cubic spline with 15 degrees of freedom



(d) PP model - No smoothing applied (e) PP model - Cubic spline with 35 degrees of freedom (f) PP model - Cubic spline with 15 degrees of freedom

Figure 3.31 Estimation of the effective reproduction number $R_e(t)$ with 50% Cr.I. for Greece based on observing deaths - multi-stage approach.

phase $T_i, i = 1, ..., K$ are distributed as $\Lambda^{-1}(E + \Lambda(T_{i-1}))$ with E drawn from Exponential(1) (Devroye, 1986).

We will use the stick-breaking representation for our model:

$$
\begin{aligned}
R_t &= r_{z_t} \\
r_j &\sim f(\cdot), \quad supp(f) = (0, \infty), \quad j = 1, ..., K \\
z_t &\sim \text{Categorical}\,(\pi_{1:K}), \quad t = 1, ..., T \\
\pi_K &= 1 - \sum_{k=1}^{K} \pi_k \\
\pi_k &= \frac{T_k}{N}, \quad k = 1, ..., K-1 \\
K &= min\{j : \sum_{i=1}^{j} T_i \geq N\} \\
T_i &= \Lambda^{-1}(E_i + \Lambda(T_{i-1})) \\
E_i &\sim \text{Exponential}\,(1), \quad i = 1, ..., K_{max} \\
\mu &\sim \text{Gamma}\,(1, 1) \\
v &\sim \text{Gamma}\,(1, 1)
\end{aligned}
\tag{3.9}
$$

setting $Kmax = 100$ as a predefined maximum number of phases.

This time-inhomogeneous Poisson process model is generally harder to estimate and can depend upon the initial values for the different MCMC chains. As a result some chains may fail to correctly estimate transmissibility (Figure 3.33). This is not entirely surprising since the K parameter is the one that is

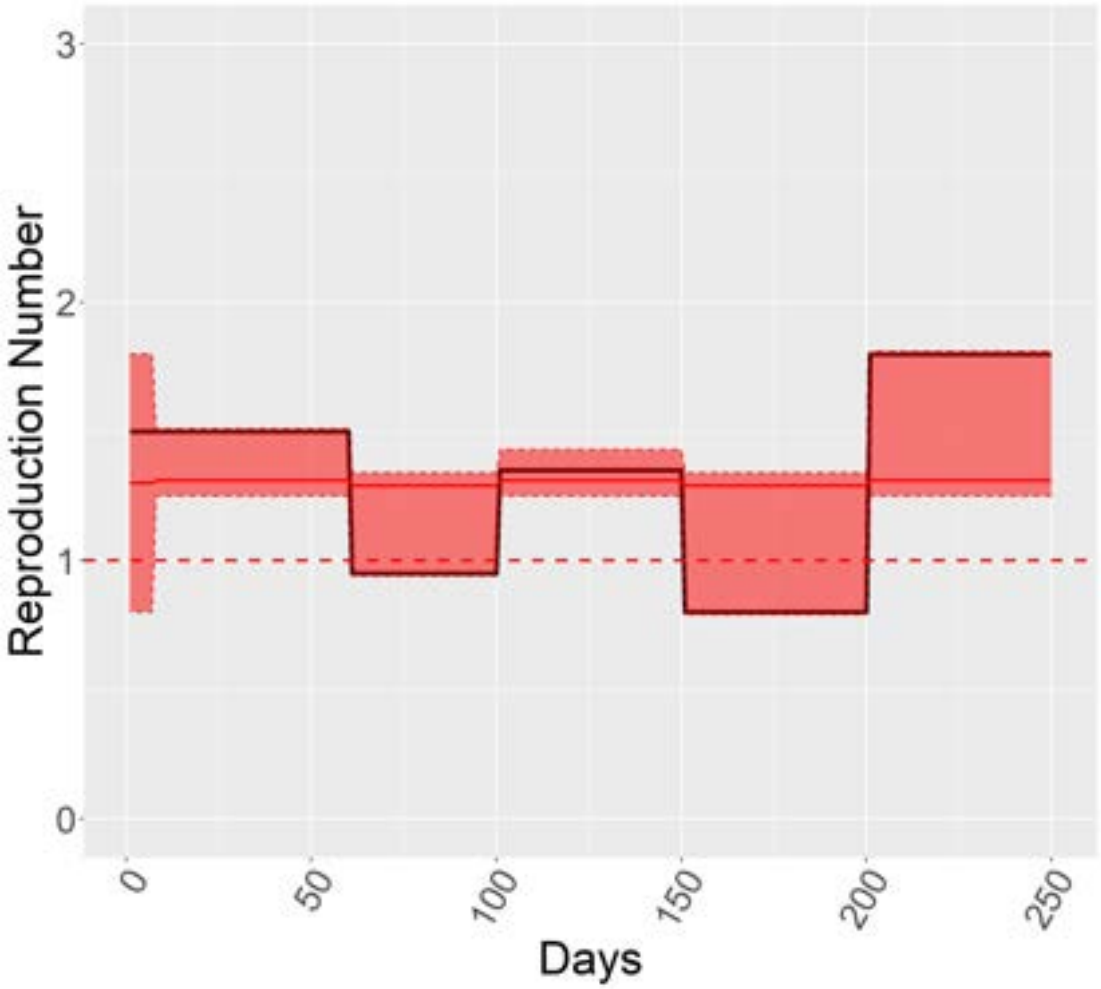


Figure 3.32 True (black line) and estimated reproduction number R_t with 95% Cr.I. (red line) based on observing infections - Non-Homogeneous Poisson process

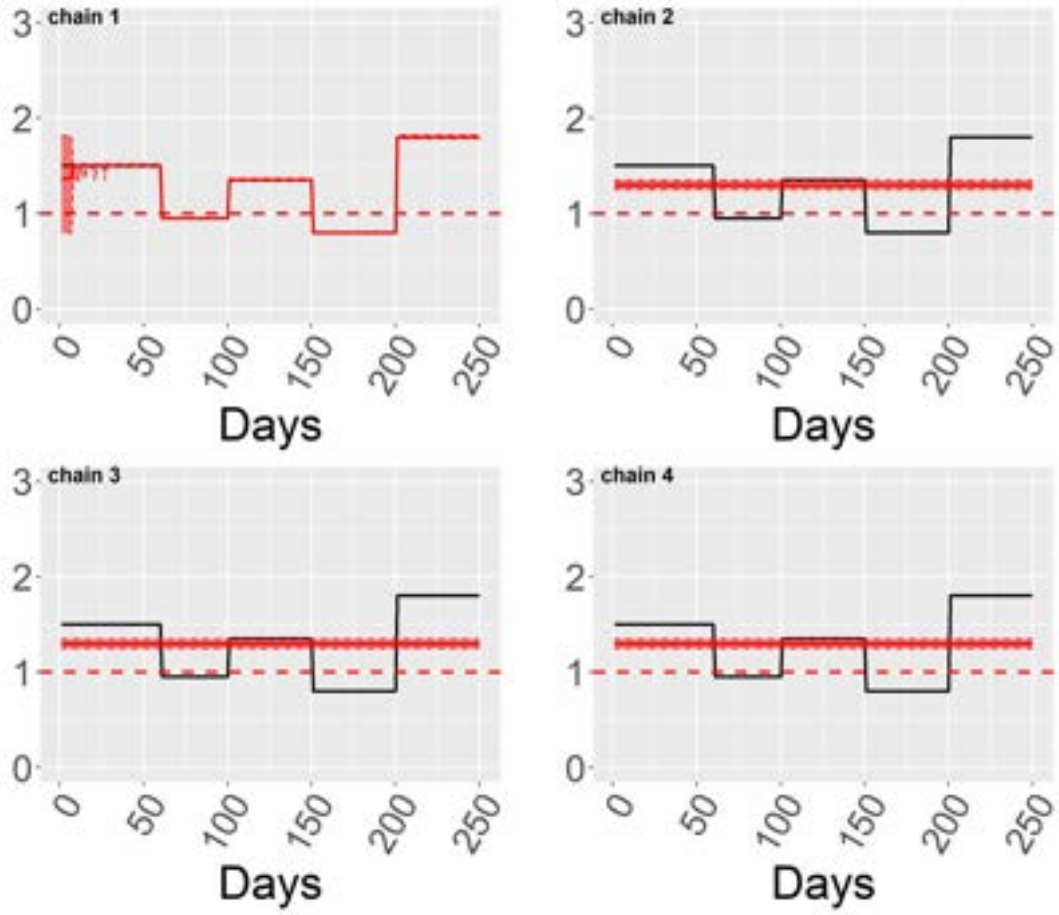


Figure 3.33 True (black line) and estimated reproduction number R_t with 95% Cr.I. (red line) based on observing infections - Non-Homogeneous Poisson process

further from the observed data. Therefore, considering the model hierarchy one may reasonably select the more parsimonious time homogeneous process as the preferred model component.

3.7.2 Two parameter Poisson-Dirichlet process

The two-parameter Poisson–Dirichlet distribution, also known as Pitman-Yor process denoted here as $PY(\alpha,\theta)$ was introduced in Pitman and Yor (1997). It can be thought of as a two-parameter extension of the Dirichlet process with $0 \leq \alpha < 1$ and $\theta > -\alpha$. Therefore $PY(0,\theta)$ reduces to the Dirichlet process with parameter θ introduced by Ferguson in Ferguson (1973).

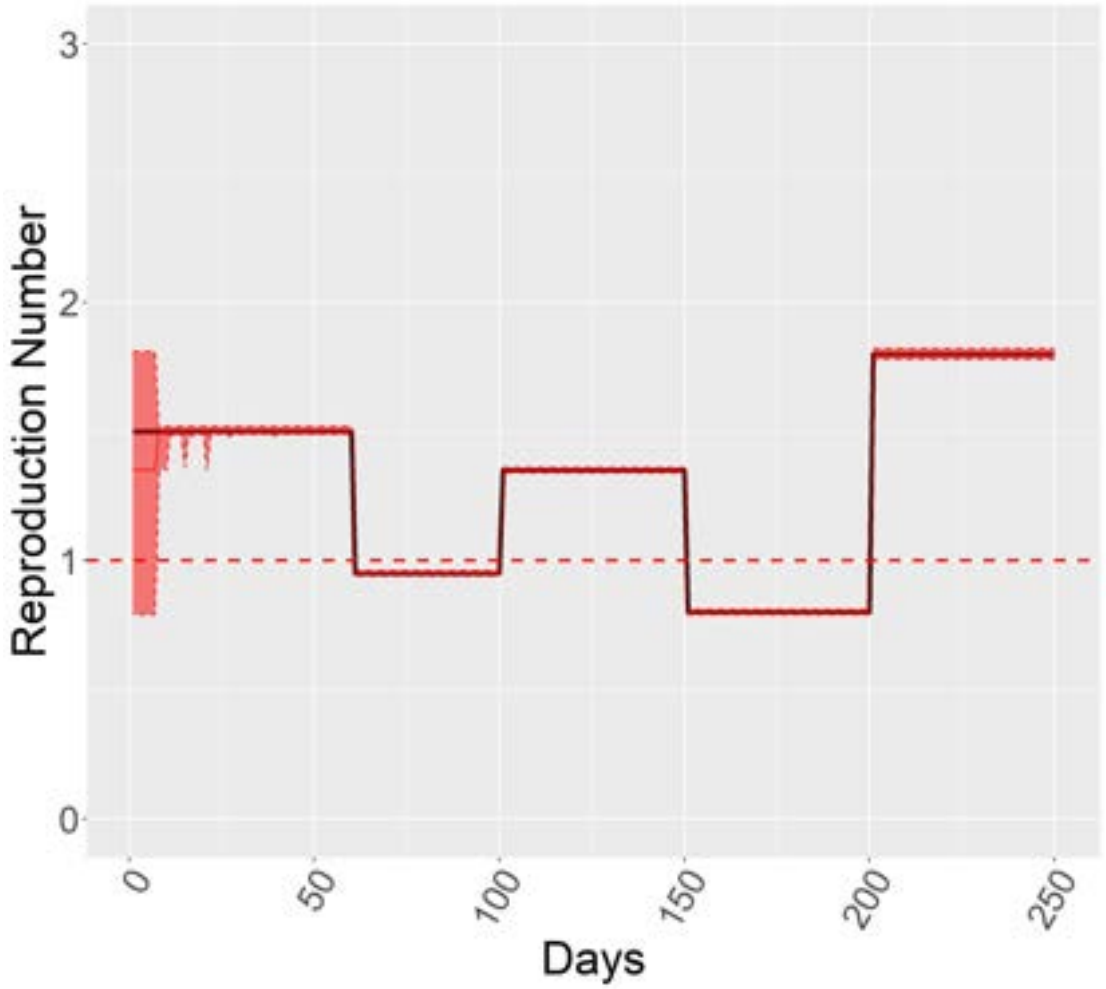


Figure 3.34 True (black line) and estimated reproduction number R_t with 95% Cr.I. (red line) based on observing infections - Pitman-Yor process

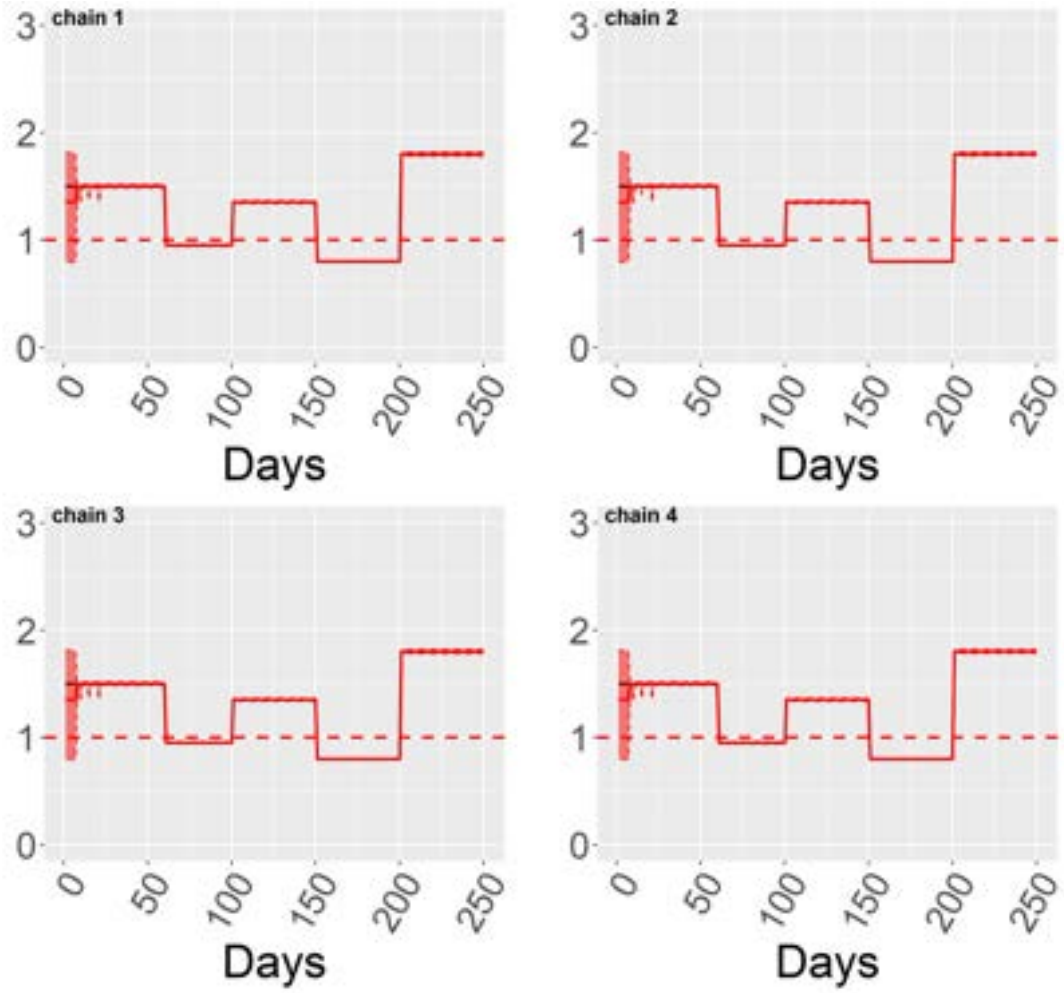


Figure 3.35 True (black line) and estimated reproduction number R_t with 95% Cr.I. (red line) based on observing infections - Pitman-Yor process

We present the stick-breaking construction of the PY:

$$
\begin{aligned}
R_t &= r_{z_t} \\
r_j &\sim f(\cdot), \quad supp(f) = (0,\infty), \quad j = 1,...,L \\
z_t &\sim \text{Categorical}\,(w_{1:L}), \quad t = 1,...,T \\
w_L &= \prod_{j<L}(1-v_j), \quad K = \sum_{j=1}^{L} I\{w_j \geq 0\} \\
w_l &= v_l * \prod_{j=1}^{l-1}(1-v_j), \quad l = 2,...,L-1 \\
w_1 &= v_1 \\
v_i &\sim \text{Beta}\,(1-\alpha, \theta + i*\alpha), \quad i = 1,...,L-1 \\
\alpha &\sim \text{Beta}\,(1,1) \\
\theta &\sim \text{Gamma}\,(1,1)
\end{aligned}
\tag{3.10}
$$

where $L = 36$ is the truncation point of the PY process.

In summary, the Pitman-Yor process gives comparable results to the Dirichlet process model. It correctly estimates the points in time when transmissibility changes, as well as the magnitude of change. Overall, we keep the DP model since it gives similar inference while it retains parsimony and it is easier to interpret.

3.8 Computation and software

We adopted the Bayesian paradigm to inference and used freely available software like *Rstan* and *Nimble* using the statistical programming language *R*. In *Rstan* we used the No-U-Turn-Sampler (NUTS) algorithm and in *Nimble* we used a mix of Random walk Metropolis-Hastings sampler and categorical sampler for the allocation parameters z_t.

The implementation of the models is possible in both software packages. We implemented the Dirichlet process (DP) and Poisson process (PP) models in *Nimble* and the fixed number of phases model in *Rstan*. The implementation of DP and PP models in *Rstan* requires marginalization of the allocation parameters since we cannot sample discrete parameters with the NUTS algorithm. We found in our simulation studies that the implementation in *Nimble* is 10-fold faster than *Rstan* for the DP and PP models for similar results, with the computational time of the DP model being lower than the PP model.

3.9 Convergence of the algorithms

Markov chain Monte Carlo algorithms for posterior sampling can suffer from convergence issues, especially for high dimensional non-linear models like the ones presented in this chapter. Here we run the models for multiple chains to check for convergence problems. In particular, we used 8 chains for both

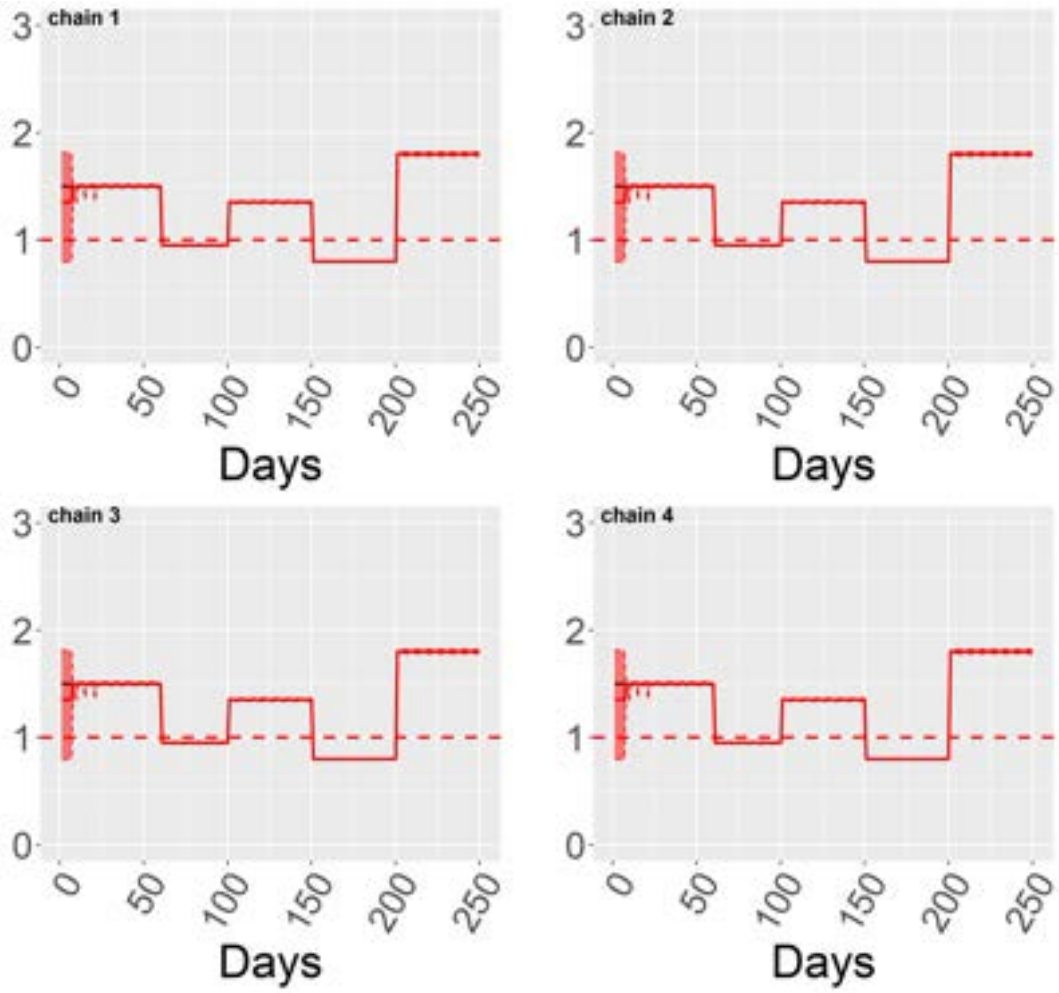


Figure 3.36 Estimation of the reproduction number R_t for the DP model for different chains of the same run based on observing infections.

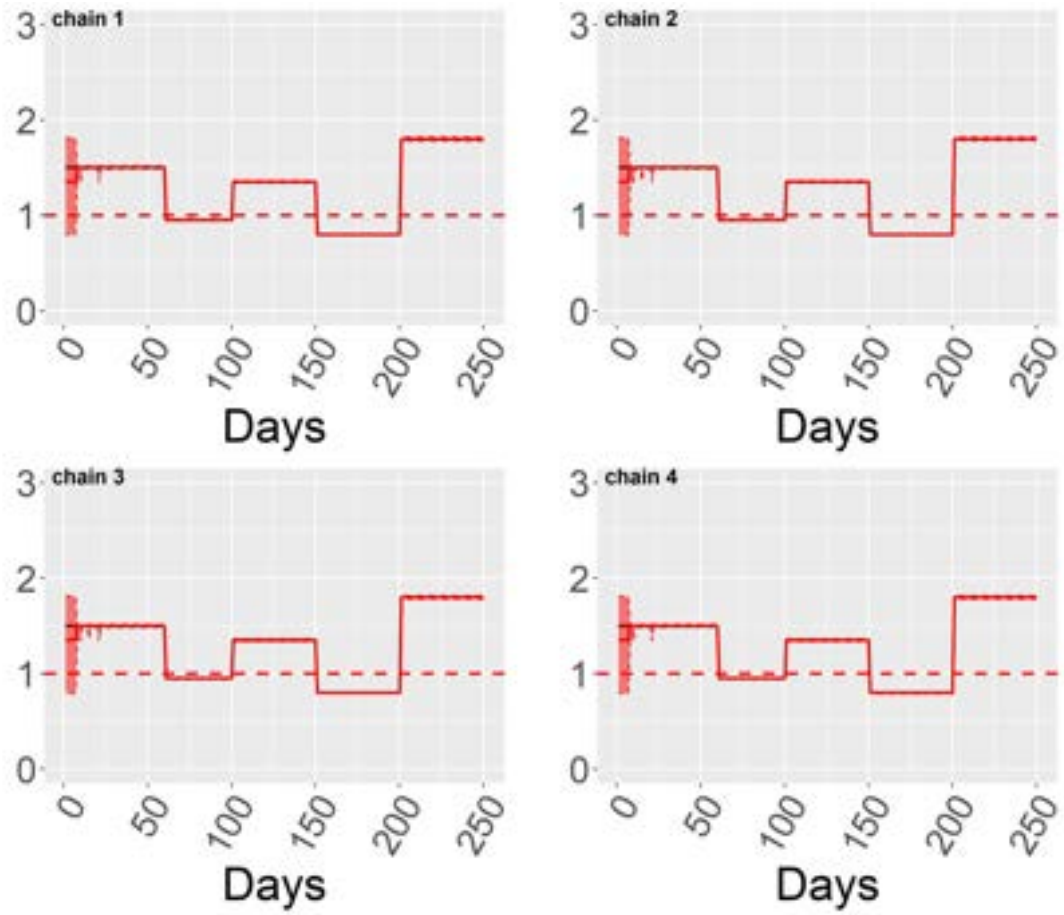
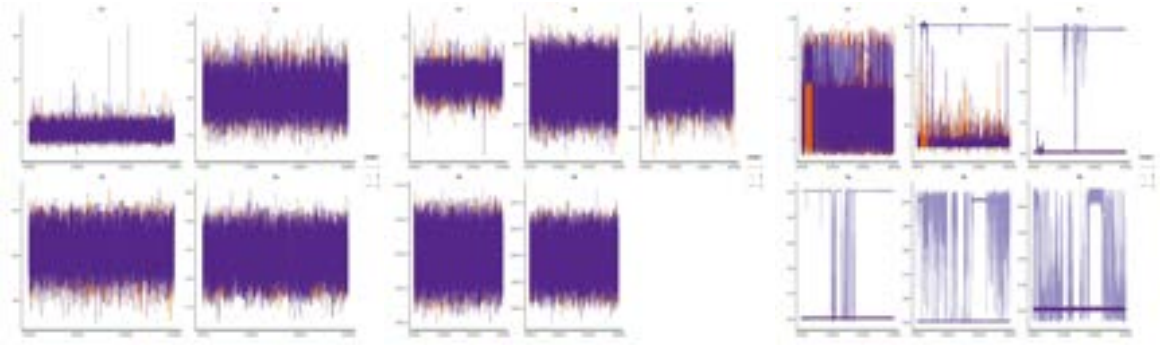


Figure 3.37 Estimation of the reproduction number R_t for the PP model for different chains of the same run based on observing infections.

models and each chain was initiated with different random values and was run for 100000 iterations using the first half of them as warm-up. Both the DP and the PP models seem robust approaches, suggesting convergence in all chains, see (Figures 3.36 and 3.37) for $R(t)$.

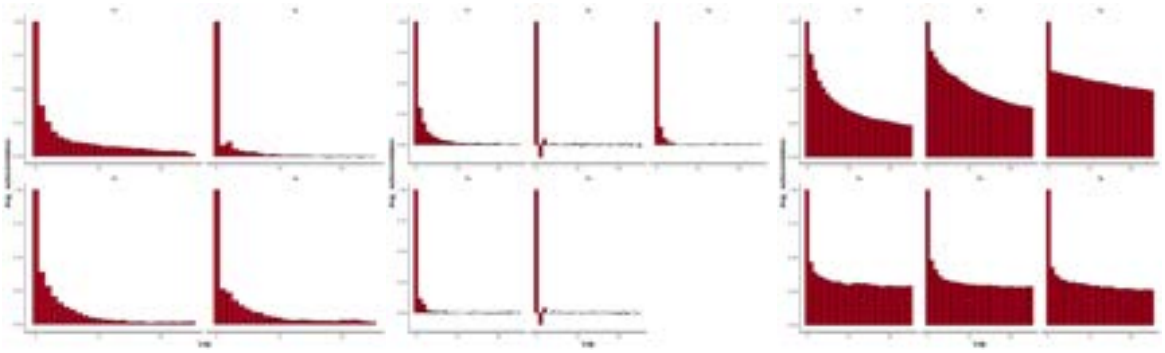


(a) 5 fixed phases (b) 6 fixed phases (c) 7 fixed phases

Figure 3.38 Trace plots for the time points of the transmissibility change - fixed number of phases model for the simulated dataset.

For the deterministic number of phases model, we checked for convergence of the Hamiltonian Monte Carlo algorithm by running 3 chains in parallel and investigating the trace and the autocorrelation plots of the returned samples. We run each chain for 30000 iterations with the first half being used as warm-up. Inspecting the trace plots of our samples (Figure 3.38) suggests that when the complexity of the model is higher than the one with the lower WAIC or LOO, there is multimodality for the points in time that the transmissibility changes. When the complexity is lower, the latter phases are entirely ignored by the model. The model with the lower WAIC or LOO values do not show

multimodality for the time that changepoints occur and have low within chain autocorrelation (Figure 3.39).



(a) 5 fixed phases (b) 6 fixed phases (c) 7 fixed phases

Figure 3.39 Autocorrelation plots for the time points of the transmissibility change - fixed number of phases model for the simulated dataset.

3.10 Discussion

In this chapter, we propose 3 models for the transmission mechanism of infectious diseases with multiple epidemic phases. We use freely available data to estimate the points in time when transmissibility changes and the realised magnitude of the NPI effects. We adopt this approach since many of these interventions coexist or overlap and identifiability issues can arise when disentangling individual effects and the associated time lags. Essentially, one may retrospectively assess the effect of the NPIs by comparing the changes in the reproduction number with the dates that these measures were imposed. Selecting the number of phases requires multiple runs and the

computation time can be an issue when nowcasting is essential for decision-making. Estimating model complexity via the DP and PP models represents an alternative approach that is computationally efficient.

The DP and PP models can estimate more epidemic phases and this issue is discussed in detail in Miller and Harrison (2013). In our setting, this effect essentially relates to the start and end of the epidemic and is inherently difficult due to limited information. At the start of the epidemic, such uncertainty dictates that estimates should be interpreted with caution. For the end this is less of an issue and is mostly due to the time lag between cases and deaths. When one is working with the observed infections these issues are largely removed and inference is typically accurate throughout the duration of the data as indicated by the simulation experiments. Our models can naturally be extended when more detailed information is available and this is the subject of further future research.

Chapter 4

Bayesian spatio-temporal regression models for the analysis of infectious diseases

4.1 Introduction

Epidemiology is the study and analysis of the distribution and factors of health and disease conditions in defined populations. Epidemic modelling is the foundation of public health and shapes policy decisions and evidence-based practice by identifying risk factors for disease transmission and allowing the assessment of preventive and curative healthcare. Epidemic models are applied for animal diseases such as Sheep-pox, and Foot and Mouth (FMD) viruses to describe outbreaks in certain countries or even in specific regions of a country. Due to the highly contagious nature of these viruses, the severe

consequences they have on animals' well-being and the significant economic consequences for people involved in animal husbandry, due to reduced milk production and weight of the livestock, it is crucial for the epidemic models to be able to predict large outbreaks to help in the virus' spread prevention.

In the literature, suitable SIR models have been used for the analysis of these types of viruses (see Jewell et al. (2009) and references within). We will use Bayesian hierarchical Poisson models embedded with stochastic processes based on stochastic differential equations similar to the work in Malesios et al. (2017, 2016). We will extend the proposed Ornstein-Uhlenbeck (OU) model by introducing a more general OU model with Student's t-distribution marginals. We will also examine the usage of the Cox-Ingersoll-Ross (CIR) model for the latent infectious rate of the model. The CIR model was introduced in 1985 by John C. Cox, Jonathan E. Ingersoll and Stephen A. Ross as an extension of the Vasicek model (Cox et al., 1985; Vasicek, 1977). The authors used the CIR model to study the term structure of interest rates which can be used in the valuation of interest rates derivatives but we think that it is underused outside the field of economics, especially in epidemic modelling. The authors in Shakiba et al. (2021) used the CIR process to model the continuous and random environmental effects on the transmission parameters of the MERS and Ebola diseases. These stochastic processes evolved around means

that are correlated with a number of covariates, including meteorological parameters and a spatial transmission kernel.

Our framework is specified in continuous time facilitating the usage of different time intervals providing flexibility based on the available data. For the Sheep-pox dataset in order to overcome the possible measurement errors of the exact infection or reporting time of disease occurrences, we will use a weekly discretization of time. The extra zeros that are generated through this discretization are suitably modelled through zero inflation probability distributions. We use contemporary Bayesian methods to perform variable selection using suitable hyper g-priors and Bayesian neural networks with horseshoe priors for the parameters of the hidden layers. For the Foot and Mouth data, we will use a daily discretization of time.

The models are applied to data from Northern Greece and especially the Northern Evros prefecture (Malesios et al., 2017, 2016). A major outbreak of sheep-pox was reported in Evros on November 1994 and lasted up until December 1998, which also resulted in several outbreaks in the neighbour prefectures and 249 infected premises. In 2000 Evros also experienced a FMD outbreak between July and September, where 5600 cattle, 4300 sheep/ goats and 360 pigs were culled during the course of the outbreak. Prevention measurements for this outbreak did not include vaccination of livestock. Farm-level data for the Sheep-pox and FMD epidemics were provided by the

Veterinary Directorate of Northern Evros Prefecture (VDNEP). Sheep-pox and FMD are widely distributed in Asia and North and East Africa, while most of Europe and America are virus free. However, the outbreaks were possible in Evros due to its unique geological position. It is a natural passage between Asia and Europe and there is a constant movement of livestock and people between Greece and Turkey. The above result in fertile ground for the transmission of diseases endemic in Western Asia and the Middle East.

4.2 Modelling epidemic data

4.2.1 Zero-inflation model

Zero inflation models are models based on zero-inflated distributions, i.e. distributions that account for frequent zero-valued data and they have been widely used in the analysis of problems with excess zero data. Zero-inflated models for count data, i.e. data $\in \{0,1,2...\}$, are mainly mixture models of Poisson or Negative-Binomial distributions. Typical use cases of these models are the modelling of the number of visits a patient makes to the emergency room in one year or the number of fish caught in one day in one lake (Bilder and Loughin, 2014). Other examples of count data are the number of hits recorded by a Geiger counter in one minute, the length of hospital stays of patients in days, goals scored in a soccer game, (Hilbe, 2011)

and the number of hypoglycemic episodes per year for a patient suffering with diabetes (Lachin, 2000).

In epidemic modelling, especially, zero-inflation models have been used for diseases that experience seasonality in their epidemic and endemic phases. It is common for these diseases to have zero new incidences for many months and then a new epidemic to evolve. Typical models without the ability to account for the excess zeros fail to catch these peaks of transmission, which is the most crucial part of disease monitoring and surveillance and guidelines of mitigation strategies imposed by health authorities. Furthermore, the information provided on the extra zeros can be useful in estimating the parameters that generate disease-free environments. In our methodology, we link environmental and spatial information on the probability of zero incidences. Zero-inflated models for count data are mainly mixture models of Poisson or Negative-Binomial distributions.

We opt to use a zero-inflated Poisson process in the spirit of Malesios et al. (2017) where the log rate of Poisson distribution is modelled through various stochastic differential equations such as Ornstein-Uhlenbeck-type processes and Cox-Inglesson-Ross model. We also use Bayesian neural nets to associate different covariates with the rate of the Poisson point process or the zero-inflation probability. The Poisson processes, where the rate is also a stochastic process, are called Cox processes (Cox, 1955). A subclass of

Cox processes where the log rate is modelled through a Gaussian process is the Log-Gaussian-Cox-Processes (LGCPs) and have been primarily used in the modelling of spatial and spatio-temporal count data of infectious diseases (Diggle et al., 2013). The authors in Malesios et al. (2017) used the aggregated spatial information through the use of kernel functions in a time-series regression framework. A different approach was used in (Diggle et al., 2013), where the lattice data of new disease incidences of the disease with prespecified positions in a $\mathbb{R}^2$ grid are modelled through LGCPs with the spatial connection between two different regions being modelled through its covariance function. We opted to use the former approach to include a spatial component in our models since in non-Gaussian cases the covariance matrix can be difficult to determine.

The likelihood of the model for modelling disease occurrences y_i at time i is as follows:

$$\begin{aligned} y_i &\sim g(y_i|\Lambda_i, p_i) \\ g(y_i|\Lambda_i, p_i) &= p_i I_{\{y_i=0\}} + (1-p_i) f(y_i|\Lambda_i) \end{aligned} \tag{4.1}$$

where $f(\cdot)$ is the probability mass function of the Poisson distribution and Λ_i is the rate.

4.2.2 Infection rate modelling

The Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930) is a stochastic process used originally in physics as a model for the velocity of a massive Brownian particle under the influence of friction. It is also used in financial mathematics and modelling, especially in the Vasicek model of the interest rate (Vasicek, 1977). A stochastic process $Z(t)$ is said to be of OU type if it satisfies a stochastic differential equation (SDE) of the form:

$$dZ_t = -\phi Z_t dt + dY(t) \tag{4.2}$$

where $Y(t)$ is the background driving Lévy process (BDLP).

Gaussian Ornstein-Uhlenbeck

We will assume that the log rate of the Poisson distribution follows an OU process with an additional drift term and the BDLP being the Brownian motion.

$$\begin{aligned} \Lambda_i &= \int_{t-1}^{t} exp(\lambda_s) ds, \quad i = 1, ..., T \\ d\lambda_t &= \phi(\lambda_t - \mu_t) + \sigma dW_t \end{aligned} \tag{4.3}$$

where W_t is the standard Wiener process, i.e. the Brownian motion, ϕ is the drift rate of the differential equation and μ is the mean reversion level. How

strong the influence of W_t on the stochastic differential equation is indicated by the term σ.

Solution of the differential equation: The differential equation in 4.3 can be explicitly solved, even if it cannot be represented integral-free. Multiplying with $e^{\phi t}$ the differential equation, and using the lemma of Itō (Itô, 1944) and the chain rule of differential calculus, we arrive at the solution $\lambda_t = \lambda_0 e^{-\phi t} + \mu_t(1 - e^{-\phi t}) + \sigma \int_0^t e^{-\phi(t-s)} dW_s$.

The conditional expectation is $E[\lambda_t|\lambda_0] = \mu_t + (\lambda_0 - \mu_t)e^{-\phi t}$.

The conditional variance is $Var[\lambda_t|\lambda_0] = \frac{\sigma^2}{2\phi}(1 - e^{-2\phi t})$.

Thus the transition density of λ_t is:

$$\lambda_{t+1}|\lambda_t \sim N(\mu_{t+1} + (\lambda_t - \mu_t)e^{-\phi}, \frac{\sigma^2}{2\phi}(1 - e^{-2\phi})) \tag{4.4}$$

Ornstein-Uhlenbeck with Student's t-distribution marginals

We extend the previous framework by assigning a suitable prior on the variance of the transition densities to obtain Student's t transition density distributions. The Student's t- distribution is symmetric and ball-shaped like the normal distribution but with heavier tails, thus allowing us to model parameters that are further from their mean, being more robust with possible outliers. The Student's t-distribution with ν degrees of freedom includes a wide range of symmetric probability distributions with power tails, ranging from the Cauchy distribution for $\nu = 1$ and the Normal distribution for $\nu \to \infty$. The

probability distribution function of the symmetric Student's t- distribution T with ν degrees of freedom :

$$f_T(x) = \frac{\Gamma(\frac{1}{2}(n+1))}{\delta\sqrt{\pi}\Gamma(\frac{1}{2}n)[1+((x-\tilde{\mu})/\delta)^2]^{(\nu+1)/2}} \tag{4.5}$$

where $\delta > 0$ is a scaling parameter and $\tilde{\mu} \in \mathbb{R}$ is a location parameter.

$E[T] = \tilde{\mu}$ for $\nu > 1$ and $Var[T] = \delta^2 \frac{\nu}{\nu-2}$ for $\nu > 2$.

A random variable $X \sim T(\nu, \tilde{\mu}, \delta)$ can be represented as $T \overset{D}{=} \tilde{\mu} + \tilde{\sigma}\varepsilon$ where the independent random variables ε and $\tilde{\sigma^2}$ have the standard normal distribution $N(0,1)$ and the inverse (reciprocal) gamma distribution $R\Gamma(\frac{1}{2}\nu, \frac{1}{2}\delta^2)$, respectively (Heyde and Leonenko, 2005; West, 1987).

The transition density of λ_t is:

$$\begin{aligned} \lambda_{t+1}|\lambda_t &\sim T(\nu, \tilde{\mu}_{t+1}, \delta) \\ \tilde{\mu}_{t+1} &= \mu_{t+1} + (\lambda_t - \mu_t)e^{-\phi} \\ \delta^2 &= \frac{\sigma^2}{2\phi}(1-e^{-2\phi}) \end{aligned} \tag{4.6}$$

The degrees of freedom ν are assigned an non-informative prior Gamma $(0.2, 0.1)$.

This process is different in construction than the Student's t-OU process presented in Heyde and Leonenko (2005). The authors used a non Brownian

motion BDLP $Y(t)$ in equation 4.2 with cumulant transform

$$\kappa_{Y(1)}(\zeta) = \log E\{exp[i\zeta Y(1)]\}, \quad \zeta \in \mathbb{R}, \quad \zeta \neq 0 \tag{4.7}$$

such that Z_t satisfies the SDE (4.2) for all $\phi > 0$ and has the solution:

$$Z_t = e^{-\phi t} X_0 + e^{-\phi t} \int_0^t e^{\phi s} dY(\phi s) \tag{4.8}$$

then Z_t has marginal t-distribution $T(\nu, \delta, \mu)$. We think our approach is more intuitive in conception and construction without the need to calculate the BDLP.

Cox–Ingersoll–Ross model

The CIR model specifies that the rate λ_t follows the following stochastic differential equation:

$$\begin{aligned} \Lambda_i &= \int_{t-1}^{t} \lambda_s ds, \quad i = 1, ..., T \\ d\lambda_t &= \alpha(e^{\mu_t} - \lambda_t) + \sigma\sqrt{\lambda_t} dW_t \end{aligned} \tag{4.9}$$

where W_t is the standard Wiener process, α is the speed to adjustment to mean level e^{μ_t} and σ corresponds to volatility. The drift factor, $\alpha(e^{\mu_t} - \lambda_t)$ is exactly the same as in the OU model. It ensures ultimately the mean reversion of the rate λ_t towards the value e^{μ_t}, with the speed of adjustment controlled by the strictly positive parameter α. The standard deviation factor $\sigma\sqrt{\lambda_t}$ ensures

the rate $\lambda_t > 0$ and for this reason, we did not transform it in the exponential scale in equation 4.9, contrary to the approach in the OU models (equation 4.3), but instead we exponentiate the term μ_t, which is the regression to covariates to keep the notation consistent. A zero value is prevented if the condition $2\alpha e^{\mu_t} \geq \sigma^2$ is met. When $_t$ becomes really small the term $\sigma\sqrt{\lambda_t}$ also becomes small neutralizing the effect of the randomness of the Wiener process and the process λ_t is guided towards equilibrium by the drift rate α.

The transition density of λ_t is:

$$\lambda_{t+T}|\lambda_t = \frac{Y_t}{2c} \tag{4.10}$$

where Y_t follows a non-central chi-square distribution with $\frac{4\alpha e^{\mu_t}}{\sigma^2}$ degrees of freedom and non-centrality parameter $2c\lambda_t e^{-\alpha T}$ The probability density function is:

$$f(\lambda_{t+T}|\lambda_t, a, \mu_t, c) = ce^{-c(\lambda_t e^{-\alpha T}+\lambda_{t+T})}\frac{\lambda_{t+T}}{\lambda_t e^{-\alpha T}}I_q(2c\sqrt{\lambda_{t+T}\lambda_t e^{-\alpha T}}) \tag{4.11}$$

where $I_q(\cdot)$ is a modified Bessel function of the first kind of order q

Due to mean reversion as $t \to \infty$, the distribution of λ_∞ approaches a Gamma distribution with shape equal to $\frac{2\alpha e^{\mu_\infty}}{\sigma^2}$ and rate equal to $\frac{2\alpha}{\sigma^2}$. The CIR model presented above with zero-inflated Poisson process likelihood can be seen as an approximated zero-inflated negative-binomial process. A point

process is a negative binomial process if it is a mixed Poisson process with intensity parameter following a Gamma distribution.

In order to simulate from the non-central chi-square distribution we use the property:

$$J \sim \text{Poisson}(\frac{1}{2}\tilde{\lambda}), \quad \text{then} \quad \chi^2_{k+2J} \sim \chi'^2_k(\tilde{\lambda}) \tag{4.12}$$

where χ^2_{k+2J} is the standard chi-square distribution with $k+2J$ degrees of freedom and $\chi'^2_k(\tilde{\lambda})$ is the non-central chi-square distribution with k degrees of freedom and non-centrality parameter $\tilde{\lambda}$.

4.2.3 Association with meteorological parameters and spatial information

Linear predictor

$$\begin{aligned} \mu_t &= X_t\beta + b_y + K(d_t, \Theta_k) \\ \text{logit}(p_t) &= X_t\beta^z + b^z_y + K(d_t, \Theta^z_k) \end{aligned} \tag{4.13}$$

where the design matrix X_t is the standardized covariate matrix containing the number of villages infected the previous week, rain, average temperature, minimum temperature, maximum temperature, humidity, wind and soil temperature of the previous week. It also contains information about the number of sheep and cows in each farm. The terms $b_y, \quad y = 1,...,5$ are random

effects for each year. The first column X_0 is assigned a unit vector and β_0 and β_0^z are the intercepts.

The vector β is the vector of the coefficients for the design matrix and is assigned a g-prior $f(\beta_{-0}|\beta_0) \sim \text{Normal}(0, \frac{ge^{\beta_0}}{n}(X_{-0}^T X_{-0})^{-1})$ with $f(\beta_0) \sim \text{Normal}(0, 10^4)$ and $\frac{g}{1+g} \sim \text{Beta}(1,1)$.

Similarly, for the zero-inflation part, the vector of the coefficients β^z is assigned a g-prior $f(\beta_{-0}^z|\beta_0^z) \sim \text{Normal}(0, \frac{g^z e^{\beta_0}}{n(1+e^{\beta_0})^2}(X_{-0}^T X_{-0})^{-1})$ with $f(\beta_0) \sim \text{Normal}(0, 10^4)$ and $\frac{g^z}{1+g^z} \sim \text{Beta}(1,1)$.

The term $K(d_i, \Theta_k)$ is an infection kernel used to model the spatial component of disease transmission, where $d_i = \{d_{kl} : k \in S_i, l \in I_{i-j}\}$, is the set of all Euclidean distances between uninfected farms $k \in S_i$ at time i and previously infected farms $l \in I_{i-j}$ within the typical exposed and infectious time of the disease. The sets S_i and I_i denote the sets of the susceptible and infectious farms at time i. This spatial kernel reflects the total burden of the disease accounting for the position between probable pairs of infector-infectee and the chance of transmission based on a function of their in-between distance. The same kernel function was used for the zero-inflation probability with a different set of parameters Θ_k^z. We use a parametric function, assuming a 3-week expose and infectious window, of the form:

$$K(d_i,\Theta_k)=\begin{cases}\frac{1}{|d_i|}\sum_{k\in S_i}\sum_{l\in I_{i-j}} bexp\{(-\frac{d_{kl}}{a})^c\}, & \text{if at least one } y_{i-j}>0, \text{ (j = 1, 2, 3)}\\ bexp\{(-\frac{d_{min}}{a})^c\}, & \text{if all } y_{i-j}=0, \text{ (j = 1, 2, 3)}\end{cases} \tag{4.14}$$

where $|d_i|$ is the cardinality of d_i and d_{min} is the minimum distance beyond which transmission cannot occur set at $250km$.

4.2.4 Bayesian Neural Network

We will use a Bayesian neural net (BNN) with one hidden layer and a hyperbolic tangent activation function. The weights on the hidden layer and the output weights will follow a horseshoe prior distribution. We will check the best model fit for a varying number of nodes in the hidden layer based on WAIC. The BNN is applied in the mean of the Poisson or in the zero-inflated probability. When the BNN is applied in the mean the zero-inflation probability is a linear regression with the covariates and the kernel (BNNLR model). When the BNN is applied to the zero-inflation probability, the log-mean of the Poisson follows a Gaussian OU, where its mean is a linear regression with the covariates and the kernel (BNNZIP model). As the number of hidden nodes increases, the prior over the function described by the neural network converges to a Gaussian process.

Below we present first the formulation of the BNN for the mean of the Poisson process.

$$\begin{aligned} \Lambda_i &= \int_{t-1}^{t} exp(\lambda_s)ds, \quad i = 1,...,T \\ \lambda_t &= w' hn_t + bias' \\ hn_{t,j} &= g(X_t w_j + bias_j + K(d_t, \Theta_k)), \quad j = 1,...,nn \end{aligned} \tag{4.15}$$

where $w' = \{w'_1,...,w'_{nn}\}$ is the vector containing the weights of the output layer, nn is the number of nodes in the hidden layer, $w_j = \{w_{j1},...,w_{jnc}\}$ is the vector of weights of that correspond to hidden node j and nc is the number of columns of the design matrix X. The hyperbolic tangent activation function: $g(x) = \frac{e^{2x}-1}{e^{2x}+1}$

Figure 4.1 presents the neural network structure graphically.

The weights of the hidden and output layers are assigned horseshoe priors, exchangeable between the weights of each layer.

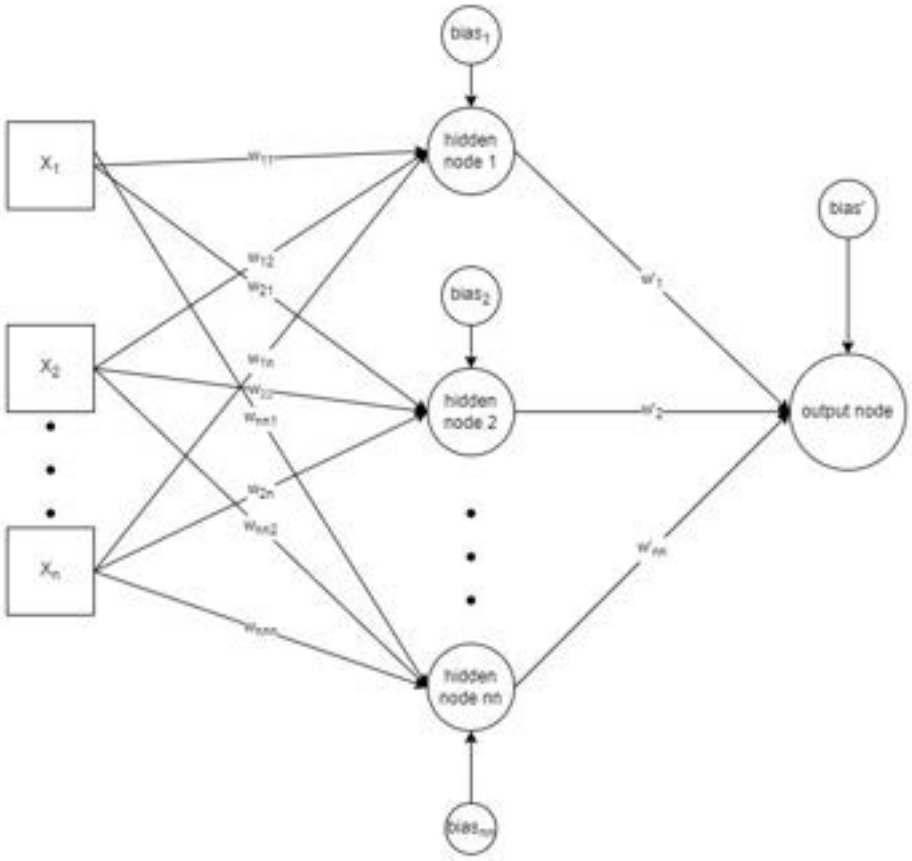


Figure 4.1 Directed acyclic graph of the Bayesian Neural Network model. Ellipses denote parameters to be learned by the model. Rectangles denote data.

$$
\begin{aligned}
w'_j &\sim \text{Normal}(0, k'_j \tau'), \quad j = 1, ..., nn \\
bias' &\sim \text{Normal}(0, k'_{bias} \tau') \\
k'_j &\sim \text{Cauchy}(0,1), \quad k'_j > 0 \\
k'_{bias} &\sim \text{Cauchy}(0,1), \quad k'_{bias} > 0 \\
\tau' &\sim \text{Gamma}(1,1) \\
w_{jl} &\sim \text{Normal}(0, k_{jl} \tau), \quad l = 1, ..., 6 \\
bias_j &\sim \text{Normal}(0, k_{biasj} \tau) \\
k_{jl} &\sim \text{Cauchy}(0,1), \quad k_{jl} > 0 \\
k_{biasj} &\sim \text{Cauchy}(0,1), \quad k_{biasj} > 0
\end{aligned}
\tag{4.16}
$$

where k and k' are the local shrinkage parameters, and t and t' are the global ones.

A similar structure with horseshoe prior is also chosen when the BNN is applied to the zero-inflation probability.

$$\begin{aligned} \text{logit}(p_t) &= w^{z\prime} hn_t^z + bias^{z\prime} \\ hn_{t,j}^z &= g(X_t w_j^z + bias_j^z + K(d_t, \Theta_k)), \quad j = 1, ..., nn \end{aligned} \tag{4.17}$$

The weights of the hidden and output layers are assigned horseshoe priors, exchangeable between the weights of each layer.

$$\begin{aligned} w^{z\prime}{}_j &\sim \text{Normal}(0, k^{z\prime}{}_j \tau^{z\prime}), \quad j = 1, ..., nn \\ bias^{z\prime} &\sim \text{Normal}(0, k'_{bias}{}^z \tau') \\ k^{z\prime}{}_j &\sim \text{Cauchy}(0,1), \quad k^{z\prime}{}_j > 0 \\ k^{z\prime}{}_{bias} &\sim \text{Cauchy}(0,1), \quad k^{z\prime}{}_{bias} > 0 \\ \tau^{z\prime} &\sim \text{Gamma}(1,1) \\ w_{jl}^z &\sim \text{Normal}(0, k_{jl}^z \tau^z), \quad l = 1, ..., 6 \\ bias_j^z &\sim \text{Normal}(0, k_{bias\,j}^z \tau) \\ k_{jl}^z &\sim \text{Cauchy}(0,1), \quad k_{jl}^z > 0 \\ k_{bias\,j}^z &\sim \text{Cauchy}(0,1), \quad k_{bias\,j}^z > 0 \\ \tau^z &\sim \text{Gamma}(1,1) \end{aligned} \tag{4.18}$$

where k^z and $k^{z\prime}$ are the local shrinkage parameters, and t^z and $t^{z\prime}$ are the global ones.

4.3 Prequential Analysis

4.3.1 Prequential Methodology

In many areas of statistical modelling, the purpose of statistical inference is mostly concentrated on forecasts on new observations given the data already observed, rather than inference on model parameters (Dawid, 1984). The prequential methodology works on a similar premise with cross-validation, where out-of-sample predictions are used in order to access a models prediction ability or access a strategies outcome, but bases its prediction for data point y_t on all previous outcomes, rather than on all outcomes distinct from y_t. The term "prequential", derives from the words predictive and sequential, and describes a general framework for assessing and comparing the predictive performance of a forecasting system of a sequential nature like a time-series model. Prequential analysis methodologies have been extensively used in the field of meteorological forecasting and in medical diagnosis problems, a review of these fields can be found on Dawid (1992).

The use of out-of-sample predictions, contrary to the methods of information criteria, has a theoretical and a practical basis. We do not want the data point y_t to influence its own prediction and there are also cases where

y_t has not yet been observed. This is especially true in the case of epidemic modelling, where forecasts about future disease incidences guide policy decisions and evidence-based practices of health authorities. Forecasts about an epidemic progression are used in identifying risk factors and proposing mitigation strategies before the actual incidences have been observed. Due to the highly contagious nature of the viruses presented in this thesis, the severe consequences it has on animals' well-being and the significant economic consequences for people involved in animal husbandry, it is crucial for the epidemic models to be able to predict large outbreaks, before they happen, to help in the virus' spread prevention. Probabilistic predictions also have the added benefit of quantifying the level of uncertainty around the forecast (quantifying basically how much we do not know), contrary to point estimates prediction methodologies.

The prequential approach can be generally described by the following simple sequential steps:

1. Based on the $Y = (y_1, y_2, \ldots, y_n)$ observed data (with n smaller than the total number of observations), calculate a forecast from a prediction distribution $P_{n+1} = P_n(y_{n+1})$, for unobserved data point y_{n+1}, where P_n corresponds to the probability forecast distribution after training with the n data points. For deterministic scoring rules calculate the point prediction $\hat{y}_{n+1}$.

2. Next, observe value y_{n+1}.

3. Calculate the accumulated prediction error (APE) score for observation y_{n+1} based on some scoring rule between P_{n+1} (probabilistic scoring rule) or $\hat{y}_{n+1}$ (deterministic scoring rule) and y_{n+1}, denoted by $S(\cdot)$.

4. Increase n by 1 and repeat steps 1 to 3 until n reaches N, where N stands for the last data point.

5. Finally, calculate the mean of all prediction error scores of step 3, to derive the mean prediction error for, which is given by:

$$APE = \frac{1}{N-n} \sum_{i=n+1}^{N} S(P_i, y_i) \tag{4.19}$$

4.3.2 Scoring Rules

The need for the evaluation of a model's ability to accurately predict future outcomes gave rise to the theoretical development of scoring rules (Gneiting et al., 2007; Gneiting and Raftery, 2007). A scoring rule is a measure of the 'distance' between a probability forecast and the observed outcome containing information simultaneously for both the sharpness and calibration of a given model. Scoring rules framework has been used in various applications in medicine (Hilden, 2018; Spiegelhalter, 1986) and in meteorology Murphy and Winkler (1977, 1984), among others. Scoring rules are used as loss

functions with the intent to minimize or maximise depending on the way they are defined, and whether they are monotonically increasing or decreasing.

Definition We will give the formal definition of a scoring rule as it was given in Gneiting et al. (2007); Gneiting and Raftery (2007):
A scoring rule is any function $S : (\mathrm{P} \times \Omega) \to \mathbb{R} \cup \{-\infty, +\infty\}$, which is P-quasi-integrable $\forall \mathrm{P} \in P$, where P is the convex class of probability measures on (A, Ω).

The expected score under the probability measure $Q \in P$ for probabilistic forecast P can be written as:

$$S(\mathrm{P}, Q) = \int S(\mathrm{P}, \omega) dQ(\omega) \tag{4.20}$$

The scoring rule S is proper relative to P if

$$S(Q, Q) \geq S(\mathrm{P}, Q), \quad \forall \mathrm{P}, Q \in P \tag{4.21}$$

and strictly proper relative to P if it is proper and

$$S(Q, Q) = S(\mathrm{P}, Q) \iff \mathrm{P} = Q \tag{4.22}$$

Strictly proper scoring rules promote honest forecasts, meaning that the lowest score is returned when minimizing the scoring rule using as forecast function the true generator function of the outcome. Strict propriety also

ensures that both calibration and sharpness are addressed (Winkler et al., 1996). Calibration refers to the statistical consistency between the probabilistic forecasts and the observations and is a joint property of the predictive distributions and the observations. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only (Czado et al., 2009). There is no automatic or clear choice of a proper scoring rule to be superior in any given situation and it may be appropriate to use a variety of scores, to take advantage of their differing emphases and strengths.

We will now present the scoring rules for the assessment of count as they appear in Czado et al. (2009).

Logarithmic scoring rule

The Logarithmic scoring rule is defined as:

$$\text{logs}(P, y) = -\log p_y \tag{4.23}$$

This score depends on the predictive distribution P only through its probability mass function p_y that it is assigned at the out-of-sample data points. We can interpret the Logarithmic score as a 'surprisal' function (how unlikely is to observe the data points based on the predictive distribution), with the goal being to minimize the expected surprise. There is also a clear connection between this score and the predictive deviance $\text{dev}(P, y) = -2\log p_y + C$,

where C is a function of data alone and is used as a standardizing term (Czado et al., 2009; Spiegelhalter et al., 2002). The authors in Gschlöl and Czado (2007, 2008) have chosen the standardizing term equal to zero, which directly corresponds to the logarithmic score multiplied by 2. The logarithmic scoring rule ignores the topology of space and is the only coherent scoring rule under exchangeability of the data (Fong and Holmes, 2019). In our cases, the data is time-ordered and we assume dependence between them. The logarithmic score also appears in the minimum description length principle, which is a formalization of Occam's Razor in which the best hypothesis for a given set of data is the one that permits the shortest encoding of the observed data together with the prediction model. The minimum description length principle was introduced by Rissanen (1978); it is important in information theory and learning theory. The optimality of using the logarithmic score under the assumption that our candidate models contain the correct model is shown in Dawid (1992). In cases when this assumption is violated, alternative prequential criteria-scoring rules also deserve attention.

Brier score

The Brier or quadratic score is defined as:

$$\text{brs}(P,y) = \sum_{i=0}^{\infty}(o_i - p_y)^2 = 1 - 2p_y + ||p||^2 \tag{4.24}$$

where o_i equals to 1, if $y = i$ and 0 otherwise and $||p||^2 = \sum_{i=0}^{\infty} p_i$. The interpretation of the Brier score is that it is an Euclidean distance in the probability domain and takes values between 0 and 1 with lower values meaning better prediction accuracy. The Brier score was originally used for binary or categorical data (BRIER, 1950), and it was proposed for the assessment of time series count data in Wecker (1989).

Spherical scoring rule

The spherical scoring rule is defined as:

$$\text{sphs}(P, y) = -\frac{p_y}{\sqrt{||p||^2}} \tag{4.25}$$

A detailed geometric characterization of the spherical scoring rule can be found in Jose (2007).

Ranked probability score

The ranked probability score was originally introduced in Epstein (1969) and was used as a scoring system for probability forecasts in ranked categorical data. We will use the definition given in Czado et al. (2009) for count data:

$$\text{rps}(P, y) = \sum_{k=0}^{\infty} \{P_k - \mathbb{1}(k \geq y)\}^2 \tag{4.26}$$

The ranked probability score has a similar interpretation to the Brier score as is the Euclidean distance of the predictive distribution and the empirical cumulative density function of the observed data. In Gneiting and Raftery (2007) a different representation was given in terms of expectations:

$$\text{rps}(P,y) = E_P|Y-y| - \frac{1}{2}E_P|Y-Y'| \tag{4.27}$$

where Y and Y' are independent copies of the predictive distribution P. When P is a point forecast the ranked probability score reduces to the absolute error.

Squared error

A classical measure of predictive ability based on point predictions is the squared error and it is defined as:

$$\text{ses}(P,y) = (y-\mu_P)^2 \tag{4.28}$$

where μ_P is the mean of the predictive distribution. The squared error is the Euclidean distance between two points in $\mathbb{R}$ without any quantification of the uncertainty of the prediction. This scoring rule is proper but not strictly proper (Gneiting and Raftery, 2007).

Normalized squared error

The normalised squared error score is defined as:

$$\text{nses}(P,y) = \left(\frac{y-\mu_P}{\sigma_P}\right)^2 \tag{4.29}$$

where σ_P is the standard deviation of the predictive distribution. This scoring rule accounts also for the predictive uncertainty and it is prefered when we have probabilistic predictions instead of point ones. The normalized squared error score is improper since it tends to zero when the σ_P tends to infinity.

David-Sebastiani score

The David-Sebastiani scoring rule is defined as:

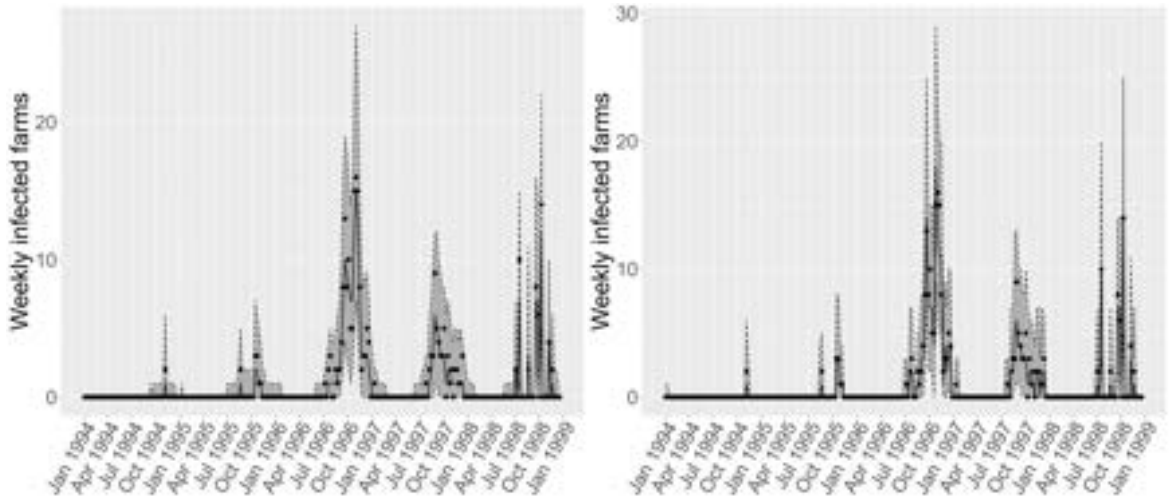
$$\text{dss}(P,y) = \left(\frac{y-\mu_P}{\sigma_P}\right)^2 + 2\log\sigma_P \tag{4.30}$$

This parametrization was proposed in Gneiting and Raftery (2007) based on works of Dawid and Sebastiani (1999) and deals with the vanishing error of the normalised squared error score when the σ_P gets sufficiently large.

4.4 Results

We trained all five models both on the Sheep-pox dataset, as well as the Foot and Mouth one. Based on the information criterion WAIC the best performing

model is the CIR for the Sheep-pox data (Table 4.1). For Foot and Mouth disease, the model with the lowest WAIC is the Poisson model with the latent rate following the Student-t OU process. All 5 models are able to estimate the trend of the infections for both datasets (Figures 4.2-4.7).



(a) Gaussian OU (b) Student-t OU

Figure 4.2 Reported and estimated weekly infections with 95% Cr.I. (solid and dashed lines), Sheep-pox data.

In order to assess the out-of-sample predictive ability of the proposed models we performed prequential analyses on both datasets. For the Sheep-pox dataset, we made out of sample predictions for the last 60 weeks of disease incidences and for the Foot and Mouth, since it is a smaller dataset we performed predictions for the last 32 days. We run each model instance for 8 chains in parallel for 200000 iterations, where the first half was used as warm-up. We used the freely available R package Nimble for the implementation of all the models, which implements Metropolis-Hastings algorithms.

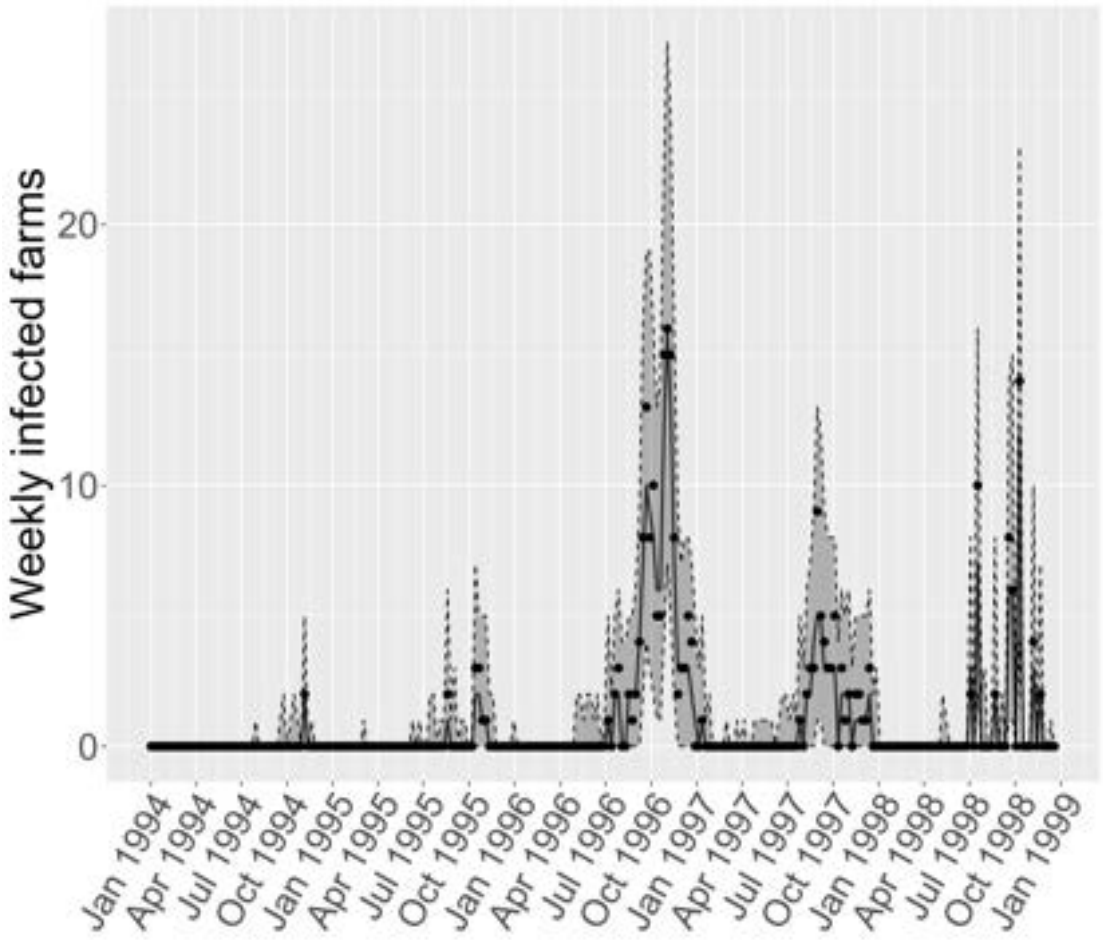
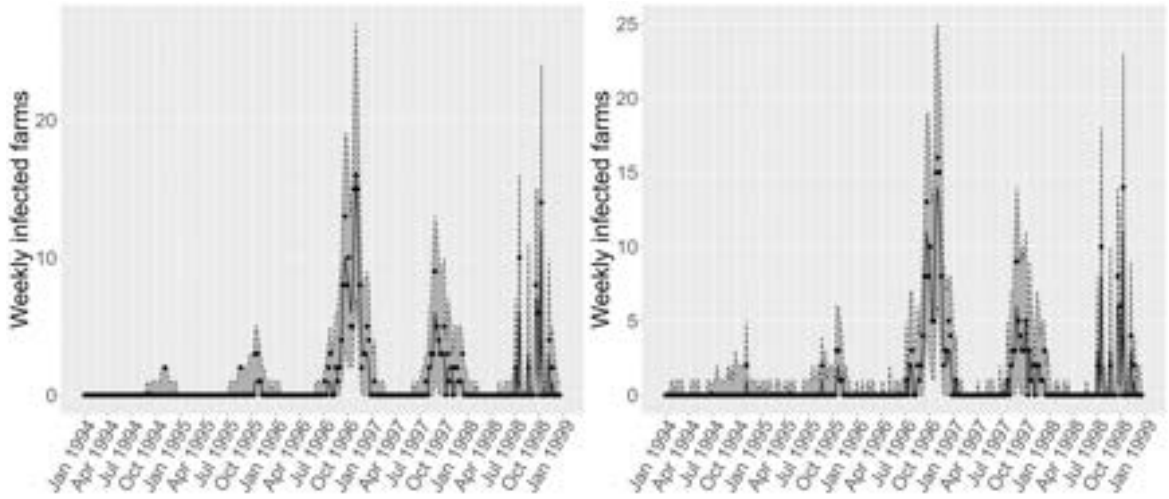


Figure 4.3 Reported and estimated weekly infections with 95% Cr.I. (solid and dashed lines), CIR model, Sheep-pox data.

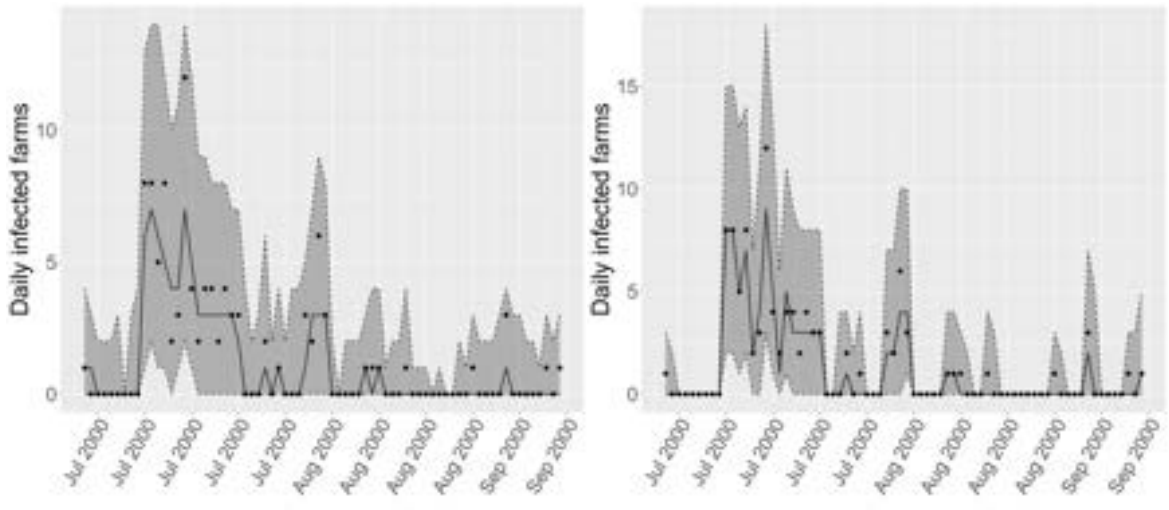
Table 4.1 WAIC for Sheep-pox dataset

Model	WAIC
Gaussian OU	289
Student OU	280
CIR	**272**
BNNZIP	294
BNNLR	296



(a) BNNZIP (b) BNNLR

Figure 4.4 Reported and estimated weekly infections with 95% Cr.I. (solid and dashed lines), Sheep-pox data.



(a) Gaussian OU (b) Student-t OU

Figure 4.5 Reported and estimated weekly infections with 95% Cr.I. (solid and dashed lines), Foot and Mouth data.

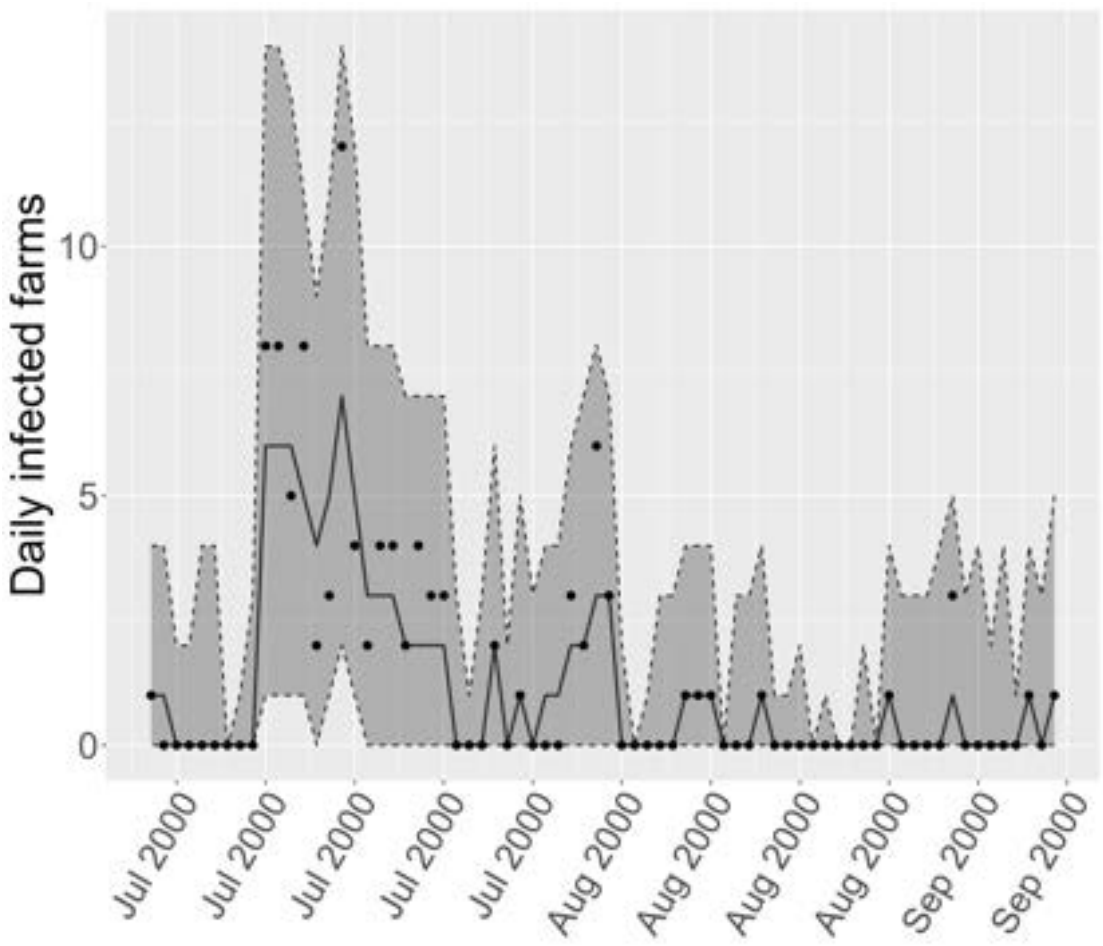
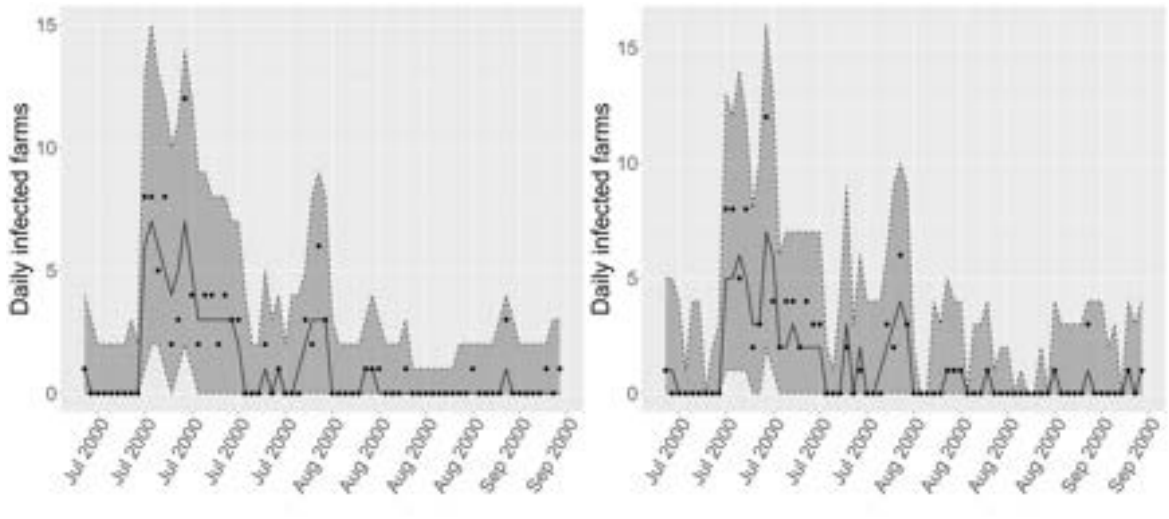


Figure 4.6 Reported and estimated weekly infections with 95% Cr.I. (solid and dashed lines), CIR model, Foot and Mouth data.

Table 4.2 WAIC for Foot and Mouth dataset

Model	WAIC
Gaussian OU	167
Student OU	**144**
CIR	162
BNNZIP	168
BNNLR	175



(a) BNNZIP (b) BNNLR

Figure 4.7 Reported and estimated weekly infections with 95% Cr.I. (solid and dashed lines), Foot and Mouth data.

In the Sheep-pox data, the Gaussian OU model has the lowest Standardize square and David and Sebastiani scores (Table 4.3). The Student-t OU performs similarly to the Gaussian OU while having the lowest Quadratic score. The main difference between the two models is observed in the Squared scoring rule, which is the only scoring rule that does not account for the variance of the predictions. The CIR model that had the lowest WAIC is not favourable by any scoring rule, although it has close Logarithmic, Ranked and Squared score values to the BNNLR, which has the lowest ones (Table 4.3). The BNNZIP presents the worst scores in this dataset.

For the Foot and Mouth dataset, the Student-t OU model has the lowest Logarithmic, Spherical, Ranked and Squared error scores. The Gaussian OU model performs identically in terms of Spherical score with the Student-t OU

Table 4.3 Scoring rules results for Sheep-pox dataset

Model	Scoring rules						
	Logarithmic	Spherical	Quadratic	Ranked	Squared	Standardized square	David & Sebastiani
Gaussian OU	4.589	-0.617	0.496	3.844	154.221	**0.678**	**0.799**
Student OU	4.814	-0.621	**0.493**	4.047	206.459	0.952	1.059
CIR	2.447	-0.604	0.553	1.461	13.034	1.235	3.744
BNNZIP	7.590	-0.6059	0.505	6.715	566.315	0.828	1.343
BNNLR	**2.187**	**-0.6363**	0.5147	**1.2985**	**10.177**	2.420	2.403

Table 4.4 Scoring rules results for Foot and Mouth dataset

Model	Scoring rules						
	Logarithmic	Spherical	Quadratic	Ranked	Squared	Standardized square	David & Sebastiani
Gaussian OU	1.316	**-0.622**	**0.573**	0.662	2.261	0.459	3.086
Student OU	**1.270**	**-0.622**	**0.574**	**0.632**	**1.997**	0.093	2.812
CIR	4.648	-0.426	0.734	3.830	85.009	**0.073**	5.449
BNNZIP	1.331	-0.591	0.627	0.672	2.152	0.265	**2.757**
BNNLR	1.563	-0.484	0.744	0.849	2.627	0.482	3.005

and has similar values on all the other scores. The CIR model presents the lowest standardized squared score, although it has the highest squared error. The BNNZIP has the lowest David & Sebastiani score(Table 4.4).

4.5 Discussion

In this Chapter, we investigated 5 different models used for modelling disease outbreaks in livestock for their predictive ability. We used Poisson process models where the latent rate is described by a series of stochastic differential equations. We generalised the Gaussian OU model presented in Malesios et al. (2017) by using its Student-t counterpart with transition densities with fatter tails, allowing theoretically for higher degrees of overdispersion and prediction of sudden outbreaks. The same is also true for the CIR model for

the rate of the point process approximating eventually a Negative Binomial process.

Both models show improvements over the Gaussian OU model, both in terms of WAIC and in the prequential analysis. For the majority of the scoring rules, the two models score better or similarly with the Gaussian OU. The main drawback of the CIR model was the higher computational time needed and for many Markov chains, the convergence was not achieved, rendering them unusable. The initial values must be properly selected to overcome this issue. The two OU models did not suffer from this problem.

The two Bayesian Neural net models did not show any improvement over the Gaussian OU one based on WAIC for the two datasets. In the Foot and Mouth dataset, the BNNZIP has a similar performance with the Gaussian OU, although slightly worse, while in the Sheep-pox dataset is the worst-performing model based on all the scoring rules. The main drawback of the two Neural net models is the loss of the interpretability of the results and the way that the disease is transmitted. Furthermore, both models also suffered from slow convergence and in some instances no convergence at all. The initial values of the parameters need to be carefully selected, something not trivial given the complicated structure of these models.

Chapter 5

Discussion

In this thesis, we developed stochastic epidemic models focused on disease outbreaks in humans, as well as livestock. We developed specific statistical methodology that can mitigate the efforts of health authorities both in epidemic surveillance through the quantification of the virus transmissibility and through strategies of optimal vaccination policies for the general population. We examined different plans of action for the vaccination regarding the effects of delayed distribution of subsequent vaccine doses for infectious diseases. These approaches can be implemented during moments of crisis, such as the Covid19 pandemic, in order to achieve incomplete herd immunity in a faster time frame.

In Chapter 1 we gave a general introduction to the main subjects of this thesis. We shared a general overview of the Bayesian methodology we followed for Chapters 3 and 4. We also presented the various algorithms

we used in order to perform Bayesian inference with an extra focus on the No-U-Turn-Sampler, an advanced variant of Hamiltonian Monte Carlo, which automatically selects an appropriate number of leapfrog steps in each iteration in order to allow the proposals to traverse the posterior without doing unnecessary work, by avoiding the random-walk behaviour that arises in random-walk Gibbs or Metropolis-Hastings samplers.

In Chapter 2 the results of a simulation-based evaluation of several policies for vaccine rollout are reported, particularly focusing on the effects of delaying the second dose of two-dose vaccines. In the presence of a limited vaccine supply, the specific policy choice is a pressing issue for several countries worldwide, and the adopted course of action will affect the extension or easing of non-pharmaceutical interventions in the next months. We employ a suitably generalised, age-structure, stochastic SEIR epidemic model that can accommodate quantitative descriptions of the major effects resulting from distinct vaccination strategies. The different rates of social contacts among distinct age groups (as well as some other model parameters) are informed by a recent survey conducted in Greece, but the conclusions are much more widely applicable. The results are summarised and evaluated in terms of the total number of deaths and infections as well as life years lost. The optimal strategy is found to be one based on fully vaccinating the elderly/at risk as quickly as possible, while extending the time interval between the two vaccine

doses to 12 weeks for all individuals below 75 years old, in agreement with epidemic theory which suggests targeting a combination of susceptibility and infectivity. This policy, which is similar to the approaches adopted in the UK and in Canada, is found to be effective in reducing deaths and life years lost in the period while vaccination is still being carried out.

In Chapter 3 we developed stochastic epidemic models suitable for estimating the disease burden and transmissibility of multiphasic epidemics. At the onset of the Covid-19 pandemic, a number of non-pharmaceutical interventions have been implemented in order to reduce transmission, thus leading to multiple phases of transmission. The disease reproduction number R_t, a way of quantifying transmissibility, has been a key part in assessing the impact of such interventions. We discuss the distinct types of transmission models used and how they are linked. We consider a hierarchical stochastic epidemic model with piece-wise constant R_t, appropriate for modelling the distinct phases of the epidemic and quantifying the true disease magnitude. The location and scale of R_t changes are inferred directly from data while the number of transmissibility phases is allowed to vary. We determine the model complexity via appropriate Poisson point process and Dirichlet process-type modelling components. The models are evaluated using synthetic data sets and the methods are applied to freely available data from California and New York states as well as the United Kingdom and Greece. We estimate the

true infected cases and the corresponding R_t, among other quantities, and independently validate the proposed approach using a large seroprevalence study. We plan to further extend our methodology by developing multi-type models, to account for a more defined structure between populations.

In Chapter 4 we focused our research on Bayesian spatio-temporal regression models for the analysis of infectious diseases. We used Poisson process or Negative Binomial models embedded with latent paths based on stochastic differential equations. We generalised previous research on this field, where the transmission parameter was based on normal Orstein-Uhlenbeck processes bu introducing OU-type models with Student's t-distribution transition densities. We also used the Cox-Ingersoll-Ross model, which is primarily used in mathematical finance, to examine if it can improve the predictive ability of our models. The means of these stochastic processes are associated with a number of covariates, including meteorological parameters and a spatial transmission kernel. We examined the usage of hyper g-priors and Bayesian neural networks with horseshoe priors for the parameters of the hidden layers for performing variable selection and identifying possible covariates that can be of use to the health authorities. A prequential analysis was performed on data from N.Evros for two types of infectious diseases, Sheep pox and Foot and mouth virus. For future research, we would like to develop a continuous time random graph model where the nodes would be spatiotemporally correlated

and the whole latent path of the infection could be inferred. Additionally, it would be particularly interesting to apply sequential monte carlo algorithms to our proposed models for online learning and compare the computational and statistical efficiency between different algorithms.

Bibliography

Amit, S., Regev-Yochay, G., Afek, A., Kreiss, Y., and Leshem, E. (2021). Early rate reductions of sars-cov-2 infection and covid-19 in bnt162b2 vaccine recipients. *The Lancet*, 397(10277):875–877.

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*. Lecture notes in statistics. Springer, New York, NY, 2000 edition.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., Gilbert, P., Janes, H., Follmann, D., Marovich, M., Mascola, J., Polakowski, L., Ledgerwood, J., Graham, B. S., Bennett, H., Pajon, R., Knightly, C., Leav, B., Deng, W., Zhou, H., Han, S., Ivarsson, M., Miller, J., and Zaks, T. (2021). Efficacy and safety of the mrna-1273 sars-cov-2 vaccine. *New England Journal of Medicine*, 384(5):403–416. PMID: 33378609.

Bernardo, J. M. and Smith, A. F. M., editors (1994). *Bayesian Theory*. John Wiley & Sons, Inc.

Betancourt, M. J. and Girolami, M. (2013). Hamiltonian monte carlo for hierarchical models.

Bhatt, S., Ferguson, N., Flaxman, S., Gandy, A., Mishra, S., and Scott, J. A. (2020). Semi-mechanistic bayesian modeling of covid-19 with renewal processes.

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S. A., Zhang, T., Gao, W., Cheng, C., Tang, X., Wu, X., Wu, Y., Sun, B., Huang, S., Sun, Y., Zhang, J., Ma, T., Lessler, J., and Feng, T. (2020). Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study. *The Lancet Infectious Diseases*, 20(8):911–919.

Bilder, C. R. and Loughin, T. M. (2014). *Analysis of Categorical Data with R*. Chapman and Hall/CRC.

Birrell, P., Blake, J., van Leeuwen, E., Gent, N., and De Angelis, D. (2021). Real-time nowcasting and forecasting of covid-19 dynamics in england: the first wave. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1829):20200279.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Bollyky, T. J. (2021). U.s. covid-19 vaccination challenges go beyond supply. *Annals of Internal Medicine*, 174(4):558–559.

BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Bubar, K. M., Reinholt, K., Kissler, S. M., Lipsitch, M., Cobey, S., Grad, Y. H., and Larremore, D. B. (2021). Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. *Science*, 371(6532):916–921.

CDC (2020). Covid-19 pandemic planning scenarios | cdc. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html. (Accessed on 02/13/2023).

Cereda, D., Manica, M., Tirani, M., Rovida, F., Demicheli, V., Ajelli, M., Poletti, P., Trentini, F., Guzzetta, G., Marziano, V., Piccarreta, R., Barone, A., Magoni, M., Deandrea, S., Diurno, G., Lombardo, M., Faccini, M., Pan, A., Bruno, R., Pariani, E., Grasselli, G., Piatti, A., Gramegna, M., Baldanti, F., Melegaro, A., and Merler, S. (2021). The early phase of the COVID-19 epidemic in lombardy, italy. *Epidemics*, 37:100528.

Champredon, D., Dushoff, J., and Earn, D. J. D. (2018). Equivalence of the erlang-distributed seir epidemic model and the renewal equation. *SIAM Journal on Applied Mathematics*, 78(6):3258–3278.

Chatzilena, A., Demiris, N., and Kalogeropoulos, K. (2022). A modelling framework for the analysis of the transmission of sars-cov2.

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am J Epidemiol*, 178(9):1505–1512.

COVID-19 Forecasting Team (2022). Variation in the COVID-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. *Lancet*, 399(10334):1469–1488.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):129–164.

Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407.

Creswell, R., Robinson, M., Gavaghan, D., Parag, K. V., Lei, C. L., and Lambert, B. (2023). A bayesian nonparametric method for detecting rapid changes in disease transmission. *Journal of Theoretical Biology*, 558:111351.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

Dawid, A. and Sebastiani, P. (1999). Coherent disperson criteria for optimal experimental design. *Annals of Statistics*, 27.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292.

Dawid, A. P. (1992). Prequential data analysis. *Lecture Notes-Monograph Series*, 17:113–126.

de Gier, B., Andeweg, S., Backer, J. A., surveillance, R. C.-., epidemiology team, Hahné, S. J., van den Hof, S., de Melker, H. E., and Knol, M. J. (2021). Vaccine effectiveness against sars-cov-2 transmission to household contacts during dominance of delta variant (b.1.617.2), the netherlands, august to september 2021. *Eurosurveillance*, 26(44).

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12(1):27–36.

Demiris, N., Kypraios, T., and Smith, L. V. (2014). On the epidemic of financial crises. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 177(3):697–723.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer New York.

Diekmann, O., Heesterbeek, H., and Britton, T. (2013). *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton University Press, Princeton.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

ELSTAT (2021). Elstat - greek statistics authority. https://www.statistics.gr/. (Accessed on 02/13/2023).

Emanuel, E. J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C., Boyle, C., Smith, M., and Phillips, J. P. (2020). Fair allocation of scarce medical resources in the time of covid-19. *New England Journal of Medicine*, 382(21):2049–2055. PMID: 32202722.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology and Climatology*, 8(6):985 – 987.

FDA (2020).

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230.

Ferreira, L. S., Canton, O., Paixão da Silva, R. L., Poloni, S., Sudbrack, V., Borges, M. E., Franco, C., Darcie Marquitti, F. M., de Moraes, J. C., de Sousa Mascena Veras, M. A., André Kraenkel, R., and Coutinho, R. M. (2021). Assessing optimal time between doses in two-dose vaccination regimen in an ongoing epidemic of sars-cov-2. *medRxiv*.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Perez-Guzman, P. N., Schmit, N., Cilloni, L., Ainslie, K. E. C., Baguelin, M., Boonyasiri, A., Boyd, O., Cattarino, L., Cooper, L. V., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Djaafara, B., Dorigatti, I., van Elsland, S. L., FitzJohn, R. G., Gaythorpe, K. A. M., Geidelberg, L., Grassly, N. C., Green, W. D., Hallett, T., Hamlet, A., Hinsley, W., Jeffrey, B., Knock, E., Laydon, D. J., Nedjati-Gilani, G., Nouvellet, P., Parag, K. V., Siveroni, I., Thompson, H. A., Verity, R., Volz, E., Walters, C. E., Wang, H., Wang, Y., Watson, O. J., Winskill, P., Xi, X., Walker, P. G. T., Ghani, A. C., Donnelly, C. A., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C., Bhatt, S.,

and Team, I. C. C.-. R. (2020). Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261.

Fong, E. and Holmes, C. (2019). On the marginal likelihood and cross-validation.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLOS ONE*, 2(8):1–12.

Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., and Hens, N. (2020). Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, march 2020. *Euro Surveill*, 25(17).

Geng, X., Katul, G. G., Gerges, F., Bou-Zeid, E., Nassif, H., and Boufadel, M. C. (2021). A kernel-modulated sir model for covid-19 contagious spread from county to continent. *Proceedings of the National Academy of Sciences*, 118(21):e2023321118.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

GOV.CA (2021). Archive 22: Recommendations on the use of covid-19 vaccines [2021-10-22] - canada.ca. https://www.canada.ca/en/public-health/services/immunization/national-advisory-committee-on-immunization-naci/recommendations-use-covid-19-vaccines.html. (Accessed on 02/13/2023).

GOV.UK (2021). Statement from the uk chief medical officers on the prioritisation of first doses of covid-19 vaccines - gov.uk. https://www.gov.uk/government/news/statement-from-the-uk-chief-medical-officers-on-the-prioritisation-of-first-doses-of-covid-19-vaccines. (Accessed on 02/13/2023).

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal - SCAND ACTUAR J*, 2007.

Gschlößl, S. and Czado, C. (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers*, 49.

Haas, E. J., Angulo, F. J., McLaughlin, J. M., Anis, E., Singer, S. R., Khan, F., Brooks, N., Smaja, M., Mircus, G., Pan, K., Southern, J., Swerdlow, D. L., Jodar, L., Levy, Y., and Alroy-Preis, S. (2021). Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in israel: an observational study using national surveillance data. *Lancet*, 397(10287):1819–1829.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

He, X., Lau, E. H. Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B. J., Li, F., and Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature Medicine*, 26(5):672–675.

Heyde, C. C. and Leonenko, N. N. (2005). Student processes. *Advances in Applied Probability*, 37(2):342–365.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press, 2 edition.

Hilden, J.; Habbema, J. D. F. B. B. (2018). Tlie measurement of performance in probabilistic diagnosis. *Methods Inf Med*, 17(04):238–246.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.

Hu, G. and Geng, J. (2021). Heterogeneity learning for sirs model: an application to the covid-19. *Statistics and Its Interface*, 14:73–81.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the dirichlet process. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(2):269–283.

Itô, K. (1944). Stochastic integral. *Proceedings of the Imperial Academy*, 20(8):519 – 524.

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4(3):465 – 496.

Jiang, F., Zhao, Z., and Shao, X. (2021). Modelling the covid-19 infection trajectory: A piecewise linear quantile trend model. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. Publisher Copyright: © 2021 Royal Statistical Society.

Jose, V. R. (2007). A characterization for the spherical scoring rule. *Theory and Decision*, 66(3):263–281.

Knock, E. S., Whittles, L. K., Lees, J. A., Perez-Guzman, P. N., Verity, R., FitzJohn, R. G., Gaythorpe, K. A. M., Imai, N., Hinsley, W., Okell, L. C., Rosello, A., Kantas, N., Walters, C. E., Bhatia, S., Watson, O. J., Whittaker, C., Cattarino, L., Boonyasiri, A., Djaafara, B. A., Fraser, K., Fu, H., Wang, H., Xi, X., Donnelly, C. A., Jauneikaite, E., Laydon, D. J., White, P. J., Ghani, A. C., Ferguson, N. M., Cori, A., and Baguelin, M. (2021). Key epidemiological drivers and impact of interventions in the 2020 sars-cov-2 epidemic in england. *Science Translational Medicine*, 13(602):eabg4262.

Koh, W. C., Naing, L., Chaw, L., Rosledzana, M. A., Alikhan, M. F., Jamaludin, S. A., Amin, F., Omar, A., Shazli, A., Griffith, M., Pastore, R., and Wong, J. (2020). What do we know about sars-cov-2 transmission? a systematic review and meta-analysis of the secondary attack rate and associated risk factors. *PLOS ONE*, 15(10):1–23.

Lachin, J. (2000). Biostatistical methods: The assessment of relative risks, second edition.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582.

Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., Rossi, L., Manganelli, R., Loregian, A., Navarin, N., Abate, D., Sciro, M., Merigliano, S., De Canale, E., Vanuzzo, M. C., Besutti,

V., Saluzzo, F., Onelia, F., Pacenti, M., Parisi, S. G., Carretta, G., Donato, D., Flor, L., Cocchio, S., Masi, G., Sperduti, A., Cattarino, L., Salvador, R., Nicoletti, M., Caldart, F., Castelli, G., Nieddu, E., Labella, B., Fava, L., Drigo, M., Gaythorpe, K. A. M., Brazzale, A. R., Toppo, S., Trevisan, M., Baldo, V., Donnelly, C. A., Ferguson, N. M., Dorigatti, I., Crisanti, A., Ainslie, K. E. C., Baguelin, M., Bhatt, S., Boonyasiri, A., Boyd, O., Coupland, H. L., Cucunubá, Z., Djafaara, B. A., van Elsland, S. L., FitzJohn, R., Flaxman, S., Green, W. D., Hallett, T., Hamlet, A., Haw, D., Imai, N., Jeffrey, B., Knock, E., Laydon, D. J., Mellan, T., Mishra, S., Nedjati-Gilani, G., Nouvellet, P., Okell, L. C., Parag, K. V., Riley, S., Thompson, H. A., Unwin, H. J. T., Verity, R., Vollmer, M. A. C., Walker, P. G. T., Walters, C. E., Wang, H., Wang, Y., Watson, O. J., Whittaker, C., Whittles, L. K., Xi, X., and Team, I. C. C.-. R. (2020). Suppression of a sars-cov-2 outbreak in the italian municipality of vo'. *Nature*, 584(7821):425–429.

Lewnard, J. A. and Lo, N. C. (2020). Scientific and ethical basis for social-distancing interventions against covid-19. *The Lancet Infectious Diseases*, 20(6):631–633.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M., and Feng, Z. (2020a). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13):1199–1207. PMID: 31995857.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020b). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490):489–493.

Li, Y. I., Turk, G., Rohrbach, P. B., Pietzonka, P., Kappler, J., Singh, R., Dolezal, J., Ekeh, T., Kikuchi, L., Peterson, J., Bolitho, A., Kobayashi, H., Cates, M. E., Adhikari, R., and Jack, R. L. (2021). Efficient bayesian inference of fully stochastic epidemiological models with applications to COVID-19. *R Soc Open Sci*, 8(8):211065.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H., Fisman, D.,

and Murray, M. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300(5627):1966–1970.

Liu, T., Hu, J., Kang, M., Lin, L., Zhong, H., Xiao, J., He, G., Song, T., Huang, Q., Rong, Z., Deng, A., Zeng, W., Tan, X., Zeng, S., Zhu, Z., Li, J., Wan, D., Lu, J., Deng, H., He, J., and Ma, W. (2020). Transmission dynamics of 2019 novel coronavirus (2019-ncov). *bioRxiv*.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., and Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359.

Malesios, C., Demiris, N., Kalogeropoulos, K., and Ntzoufras, I. (2017). Bayesian epidemic models for spatially aggregated count data. *Statistics in Medicine*, 36(20):3216–3230.

Malesios, C., Demiris, N., Kostoulas, P., Dadousis, K., Koutroumanidis, T., and Abas, Z. (2016). Spatio-temporal modelling of foot-and-mouth disease outbreaks. *Epidemiol. Infect.*, 144(12):2485–2493.

Matrajt, L., Eaton, J., Leung, T., and Brown, E. R. (2021a). Vaccine optimization for COVID-19: Who to vaccinate first? *Science Advances*, 7(6).

Matrajt, L., Eaton, J., Leung, T., Dimitrov, D., Schiffer, J. T., Swan, D. A., and Janes, H. (2021b). Optimizing vaccine allocation for covid-19 vaccines: potential role of single-dose vaccination. *medRxiv*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Miller, J. W. and Harrison, M. T. (2013). Inconsistency of Pitman-Yor process mixtures for the number of components.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356. PMID: 29983475.

Moghadas, S. M., Vilches, T. N., Zhang, K., Nourbakhsh, S., Sah, P., Fitzpatrick, M. C., and Galvani, A. P. (2021). Evaluation of covid-19 vaccination strategies with a delayed second dose. *PLOS Biology*, 19(4):1–13.

Moore, S., Hill, E. M., Tildesley, M. J., Dyson, L., and Keeling, M. J. (2021). Vaccination and non-pharmaceutical interventions for covid-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 21(6):793–802.

Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(1):41–47.

Murphy, A. H. and Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79(387):489–500.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Neal, R. (2011). MCMC using hamiltonian dynamics. In *Chapman &amp Hall/CRC Handbooks of Modern Statistical Methods*. Chapman and Hall/CRC.

Nesterov, Y. (2007). Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259.

Paltiel, A. D., Schwartz, J. L., Zheng, A., and Walensky, R. P. (2021a). Clinical outcomes of a covid-19 vaccine: Implementation over efficacy. *Health Affairs*, 40(1):42–52. PMID: 33211536.

Paltiel, A. D., Zheng, A., and Schwartz, J. L. (2021b). Speed versus efficacy: Quantifying potential tradeoffs in covid-19 vaccine deployment. *Annals of Internal Medicine*, 174(4):568–570.

Parino, F., Zino, L., Calafiore, G. C., and Rizzo, A. (2021). A model predictive control approach to optimally devise a two-dose vaccination rollout: A case study on covid-19 in italy. *International Journal of Robust and Nonlinear Control*, n/a(n/a).

Pellis, L., Birrell, P. J., Blake, J., Overton, C. E., Scarabel, F., Stage, H. B., Brooks-Pollock, E., Danon, L., Hall, I., House, T. A., Keeling, M. J., Read, J. M., and and, D. D. A. (2022). Estimation of reproduction numbers in real time: Conceptual and statistical challenges. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(S1).

Persad, G., Wertheimer, A., and Emanuel, E. J. (2009). Principles for allocation of scarce medical interventions. *The Lancet*, 373(9661):423–431.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855 – 900.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., Türeci, O., Nell, H., Schaefer, A., Ünal, S., Tresnan, D. B., Mather, S., Dormitzer, P. R., Şahin, U., Jansen, K. U., and Gruber, W. C. (2020). Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615. PMID: 33301246.

Quintana, F. A., Mueller, P., Jara, A., and MacEachern, S. N. (2020). The dependent dirichlet process and related models.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Romero-Brufau, S., Chopra, A., Ryu, A. J., Gel, E., Raskar, R., Kremers, W., Anderson, K. S., Subramanian, J., Krishnamurthy, B., Singh, A., Pasupathy, K., Dong, Y., O'Horo, J. C., Wilson, W. R., Mitchell, O., and Kingsley, T. C. (2021). Public health impact of delaying second dose of BNT162b2 or mRNA-1273 covid-19 vaccine: simulation agent based modeling study. *BMJ*, page n1087.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Shakiba, N., Edholm, C. J., Emerenini, B. O., Murillo, A. L., Peace, A., Saucedo, O., Wang, X., and Allen, L. J. (2021). Effects of environmental variability on superspreading transmission events in stochastic epidemic models. *Infectious Disease Modelling*, 6:560–583.

Shen, Z., Ning, F., Zhou, W., He, X., Lin, C., Chin, D. P., Zhu, Z., and Schuchat, A. (2004). Superspreading SARS events, beijing, 2003. *Emerg Infect Dis*, 10(2):256–260.

Silva, P. J. S., Sagastizábal, C., Nonato, L. G., Struchiner, C. J., and Pereira, T. (2021). Optimized delay of the second covid-19 vaccine dose reduces icu admissions. *Proceedings of the National Academy of Sciences*, 118(35):e2104640118.

Singanayagam, A., Hakki, S., Dunning, J., Madon, K. J., Crone, M. A., Koycheva, A., Derqui-Fernandez, N., Barnett, J. L., Whitfield, M. G., Varro, R., Charlett, A., Kundu, R., Fenn, J., Cutajar, J., Quinn, V., Conibear, E., Barclay, W., Freemont, P. S., Taylor, G. P., Ahmad, S., Zambon, M., Ferguson, N. M., Lalvani, A., Badhan, A., Dustan, S., Tejpal, C., Ketkar, A. V., Narean, J. S., Hammett, S., McDermott, E., Pillay, T., Houston, H., Luca, C., Samuel, J., Bremang, S., Evetts, S., Poh, J., Anderson, C., Jackson, D., Miah, S., Ellis, J., and Lackenby, A. (2022). Community transmission and viral load kinetics of the sars-cov-2 delta (b.1.617.2) variant in vaccinated and unvaccinated individuals in the uk: a prospective, longitudinal, cohort study. *The Lancet Infectious Diseases*, 22(2):183–195.

Skowronski, D. M. and De Serres, G. (2021). Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 384(16):1576–1578. PMID: 33596348.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5(5):421–433.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stan Development Team (2023). RStan: the R interface to Stan. R package version 2.21.8.

Sypsa, V., Roussos, S., Engeli, V., Paraskevis, D., Tsiodras, S., and Hatzakis, A. (2021a). Trends in covid-19 vaccination intent, determinants and reasons for vaccine hesitancy: results from repeated cross-sectional surveys in the adult general population of greece during november 2020-june 2021. *medRxiv*.

Sypsa, V., Roussos, S., Paraskevis, D., Lytras, T., Tsiodras, S., and Hatzakis, A. (2021b). Effects of social distancing measures during the first epidemic

wave of severe acute respiratory syndrome infection, greece. *Emerging Infectious Disease journal*, 27(2):452.

Taipale, J., Kontoyiannis, I., and Linnarsson, S. (2021). Population-scale testing can suppress the spread of infectious disease.

Thompson, M. G., Burgess, J. L., Naleway, A. L., Tyner, H. L., Yoon, S. K., Meece, J., Olsho, L. E., Caban-Martinez, A. J., Fowlkes, A., Lutrick, K., Kuntz, J. L., Dunnigan, K., Odean, M. J., Hegmann, K. T., Stefanski, E., Edwards, L. J., Schaefer-Solle, N., Grant, L., Ellingson, K., Groom, H. C., Zunie, T., Thiese, M. S., Ivacic, L., Wesley, M. G., Lamberte, J. M., Sun, X., Smith, M. E., Phillips, A. L., Groover, K. D., Yoo, Y. M., Gerald, J., Brown, R. T., Herring, M. K., Joseph, G., Beitel, S., Morrill, T. C., Mak, J., Rivers, P., Harris, K. M., Hunt, D. R., Arvay, M. L., Kutty, P., Fry, A. M., and Gaglani, M. (2021). Interim estimates of vaccine effectiveness of BNT162b2 and mRNA-1273 COVID-19 vaccines in preventing SARS-CoV-2 infection among health care personnel, first responders, and other essential and frontline workers — eight u.s. locations, december 2020–march 2021. *MMWR. Morbidity and Mortality Weekly Report*, 70(13):495–500.

Tuite, A. R., Zhu, L., Fisman, D. N., and Salomon, J. A. (2021). Alternative dose allocation strategies to increase benefits from constrained covid-19 vaccine supply. *Annals of Internal Medicine*, 174(4):570–572.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Phys. Rev.*, 36:823–841.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5(2):177–188.

Vasileiou, E., Simpson, C. R., Robertson, C., Shi, T., Kerr, S., Agrawal, U., Akbari, A., Bedston, S., Beggs, J., Bradley, D., Chuter, A., de Lusignan, S., Docherty, A., Ford, D., Hobbs, R., Joy, M., Katikireddi, S. V., Marple, J., McCowan, C., McGagh, D., McMenamin, J., Moore, E., Murray, J.-L., Pan, J., Ritchie, L., Shah, S. A., Stock, S., Torabi, F., Tsang, R. S. M., Wood, R., Woolhouse, M., and Sheikh, A. (2021). Effectiveness of first dose of COVID-19 vaccines against hospital admissions in scotland: National prospective cohort study of 5.4 million people. *SSRN Electronic Journal*.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., Angus, B., Baillie, V. L., Barnabas, S. L., Bhorat, Q. E., Bibi, S., Briner, C., Cicconi, P., Collins, A. M., Colin-Jones, R., Cutland, C. L., Darton, T. C., Dheda, K., Duncan, C. J. A., Emary, K. R. W., Ewer, K. J., Fairlie, L., Faust, S. N., Feng, S., Ferreira, D. M., Finn, A., Goodman, A. L., Green, C. M., Green, C. A., Heath, P. T., Hill, C., Hill, H., Hirsch, I., Hodgson, S. H. C., Izu, A., Jackson, S., Jenkin, D., Joe, C. C. D., Kerridge, S., Koen, A., Kwatra, G., Lazarus, R., Lawrie, A. M., Lelliott, A., Libri, V., Lillie, P. J., Mallory, R., Mendes, A. V. A., Milan, E. P., Minassian, A. M., McGregor, A., Morrison, H., Mujadidi, Y. F., Nana, A., O'Reilly, P. J., Padayachee, S. D., Pittella, A., Plested, E., Pollock, K. M., Ramasamy, M. N., Rhead, S., Schwarzbold, A. V., Singh, N., Smith, A., Song, R., Snape, M. D., Sprinz, E., Sutherland, R. K., Tarrant, R., Thomson, E. C., Török, M. E., Toshner, M., Turner, D. P. J., Vekemans, J., Villafana, T. L., Watson, M. E. E., Williams, C. J., Douglas, A. D., Hill, A. V. S., Lambe, T., Gilbert, S. C., Pollard, A. J., and Oxford COVID Vaccine Trial Group (2020). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in brazil, south africa, and the UK. *Lancet*, 397(10269):99–111.

Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.*, 160(6):509–516.

Wang, W., Wu, Q., Yang, J., Dong, K., Chen, X., Bai, X., Chen, X., Chen, Z., Viboud, C., Ajelli, M., and Yu, H. (2020). Global, regional, and national estimates of target population sizes for covid-19 vaccination: descriptive study. *BMJ*, 371.

Ward, H., Atchison, C., Whitaker, M., Ainslie, K. E. C., Elliott, J., Okell, L., Redd, R., Ashby, D., Donnelly, C. A., Barclay, W., Darzi, A., Cooke, G., Riley, S., and Elliott, P. (2021). Sars-cov-2 antibody prevalence in england following the first peak of the pandemic. *Nature Communications*, 12(1):905.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *J. Mach. Learn. Res.*, 14(1):867–897.

Wecker, W. E. (1989). Comment: Assessing the accuracy of time series mode! forecasts of count observations. *Journal of Business & Economic Statistics*, 7(4):418–419.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.

Winkler, R. L., Muñoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M., and Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60.

Wistuba, T., Mayr, A., and Staerk, C. (2022). Estimating the course of the covid-19 pandemic in germany via spline-based hierarchical modelling of death counts. *Scientific Reports*, 12(1):9784.

Åke Svensson (2007). A note on generation times in epidemic models. *Mathematical Biosciences*, 208(1):300–311.

Åke Svensson (2015). The influence of assumptions on generation time distributions in epidemic models. *Mathematical Biosciences*, 270:81–89.