UNIVERSITY OF CALIFORNIA,
IRVINE

Bayesian Nowcasting of Pathogen Transmission Dynamics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Isaac H Goldstein

Dissertation Committee:
Professor Volodymyr M Minin, Chair
Professor Veronica Berrocal
Professor Babak Shahbaba

2024

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to thank my advisor, Volodymyr Minin, for his support for nearly five years. He has gone well above and beyond the call of duty, and continues to set the very highest standard for mentorship. I feel incredibly lucky to have ended up in his lab.

Within the Minin lab, thanks to Jon Fintzi for patiently helping me at my most computer-illiterate to install computing software on my laptop, and for volunteering his time to helping me understand epidemic models early in the SARS-CoV-2 pandemic. Thanks also to Damon Bayer for on the job training in coding and visualization in the early days of the SARS-CoV-2 pandemic, and for four years of collaboration and consultation on a range of research endeavors. Any remaining imperfections in my code (and there are many) are entirely the result of my own errors. Research is a lonely pursuit, and I will be forever grateful to my lab mates–Catalina Medina, Thanasi Bakis, Jessalyn Sebastian and Christian Bernal, and especially Damon Bayer–for their good humor, support, and companionship. I couldn't have asked for better fellow travelers.

In addition, I have been blessed with exceptional collaborators from a range of disciplines, Jon Wakefield, Sanghyuk Shin, Sunny Jiang, and Daniel Parker, Tomás León and the rest of the CDPH team made collaboration easy, fun, and educational. They have my sincere thanks.

I would also like to thank my family, a constant source of support and care. Finally, I must thank my partner, Rachel. Without her, this work would simply not have been possible.

This work was supported in part by funding from UC CDPH Modeling Consortium and California Collaborative Pandemic Recovery & Readiness Research.

This work utilized the infrastructure for high-performance and high-throughput computing, research data storage and analysis, and scientific software tool integration built, operated, and updated by the Research Cyberinfrastructure Center (RCIC) at the University of California, Irvine.

Chapter 3 of the dissertation is a modified version of the material published in *Journal of the Royal Statistical Society Series A* [Goldstein et al., 2024], used with permission from Oxford Press.

VITA

Isaac H Goldstein

EDUCATION

Doctor of Philosophy in Statistics **2024**
University of California, Irvine *Irvine, CA*

Master of Science in Statistics **2021**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Mathematics **2016**
Lewis & Clark College *Portland, OR*

ABSTRACT OF THE DISSERTATION

Bayesian Nowcasting of Pathogen Transmission Dynamics

By

Isaac H Goldstein

Doctor of Philosophy in Statistics

University of California, Irvine, 2024

Professor Volodymyr M Minin, Chair

A central task in statistical analyses of infectious disease surveillance data is nowcasting transmission dynamics, understanding how transmissible a pathogen is in the present day. One way to summarize transmissibility is through the effective reproduction number, the average number of individuals an individual infected today would subsequently infect under current conditions. When the effective reproduction number is above one, an outbreak is expected to grow, the reverse is true when it is below one. Estimating the effective reproduction number from observed data is non-trivial, as epidemics are only ever partially observed, and existing data streams are subject to ascertainment biases that must be taken into account. Ideally, epidemics would be modeled as a partially observed stochastic process, but in practice this is computationally prohibitive. In this dissertation, we develop statistical models for estimating the effective reproduction number from a variety of data sources using a series of computationally tractable approximate models of epidemics. In particular, we develop models for estimating the effective reproduction number from case and test data, from pathogen genome concentrations collected from wastewater in large populations, and pathogen genome concentrations collected from wastewater in small populations. We compare our methods against state-of-the-art methods in simulation studies, and apply our methods to estimate the effective reproduction number of SARS-CoV-2 in California from 2020 to 2022.

Chapter 1

Introduction

1.1 Surveillance Data

When modeling infectious disease epidemics, statisticians usually make use of what are referred to as surveillance data. Surveillance data are collected in an ongoing process by some kind of institutional apparatus, often a public health department, as part of its mission to track the spread of infectious diseases. We could contrast surveillance data with data collected for a particular study to answer a particular scientific question of interest. Common data sources include counts of cases, counts of new deaths, or counts of those currently hospitalized who have tested positive for a disease. Each data source provides information about one particular aspect of the spread of infectious disease; for instance, case data might reflect the number of new infections, while hospitalizations represent the number of individuals in an advanced stage of infection.

A number of aspects of the data collection process can further distort data. Tests may not be 100% accurate. Policies about who can test can change over time. Capacity to test can change over time as well, as can the willingness of the population to test in the first place. A more fundamental issue with surveillance data is that surveillance data are always a proxy for unobserved events that are the actual events that dictate the course of the epidemic. That is, the data that actually describe the course of an epidemic are the exact times at which individuals became infectious, infected others, and recovered, and these quantities are never fully observed.

More recently, pathogen genome concentrations collected from wastewater (often referred to as wastewater data in this dissertation) have become widely available as a new source of data. Individuals infected with a pathogen can shed copies of the pathogen through their fecal matter. These individual copies are aggregated in the sewer system, and measures of the concentration of pathogen genomes found in wastewater can be calculated [Hillary et al., 2020, Polo et al., 2020]. The concentration, then, represents a noisy measurement of the number of currently infectious or recently recovered individuals in the population. While these data are also distorted by the dynamics of the sewer system and the measurement process [Wade et al., 2022], it largely avoids the biases of case data, because there are no issues with testing capacity or willingness to test. Incorporating wastewater data into statistical models of infectious disease outbreaks has barely begun, this dissertation contributes to that effort.

1.2 The Effective Reproduction Number

In an infectious disease epidemic, the effective reproduction number at time t (or the "instantaneous effective reproduction number" [Gostic et al., 2020]) is the average number of individuals a person infected at time t would subsequently infect, if the conditions of the epidemic at time t remained the same. It is used as a measure of how "under control" an epidemic is, if the effective reproduction number is above 1, the epidemic is expected to continue to spread, while when it is below 1, the opposite is true. This parameter is time-varying, and it is of interest both to estimate changes over time retrospectively, as well as to estimate what the effective reproduction number is in the present, in the midst of an on-going epidemic. This real-time estimation problem is often referred to as “nowcasting.” The effective reproduction number is only one summary of the state of an epidemic, and as an average obviously elides differences in individual transmission dynamics [Jewell and Lewnard, 2022]. Nevertheless, it is a useful measure of transmission dynamics, and this dissertation will be focused on developing models that estimate the effective reproduction number.

1.3 Models of Infectious Disease Epidemics

Mathematical models of infectious disease epidemics begin with assumptions on individual transmission dynamics, though the final models used for inference are often population level models where the individuals in the population follow the assumed dynamics. In one kind of model, individuals are assumed to be independent and identically distributed, leading to branching process inspired models [Fraser, 2007, Cori et al., 2013]. In branching process models, the number of newly infected individuals is assumed to be related to the product of the effective reproduction number and a weighted sum of the previously infected individuals, where the weights are chosen to reflect the infectiousness of individuals so that more recently infected individuals are the most infectious. Branching process inspired models are by far the most common type of models used when estimating the effective reproduction number.

Another approach is to assume that individuals interact with each other in the context of a population, where some individuals are susceptible to infection, and others are not. This set of assumptions leads to compartmental models, which are the traditional type of model for modeling infectious disease epidemics [Andersson and Britton, 2012, Blackwood and Childs, 2018, Allen, 2010]. In a compartmental model, the total number of individuals who are currently susceptible, infectious, and recovered are tracked over time. The effective reproduction number can also be derived from compartmental models.

There are deterministic and stochastic versions of both classes of models, and both versions are used in practice [Cori et al., 2013, Bhatt et al., 2023, Fintzi, 2018, Bayer et al., 2024]. Deterministic models are computationally faster and less prone to implementation issues, while stochastic models incorporate more of the uncertainty inherent in the epidemic process. Deterministic models perform best in large population settings, while in a small population, a stochastic model may be needed. This is mathematically justified by thinking of the deterministic model as the limit of the stochastic model when the population size diverges to infinity [Kurtz, 1971, Andersson and Britton, 2012].

1.4 Motivating Examples

1.4.1 SARS-CoV-2 Cases and Tests in Orange County, California

Chapters 3–5 of the dissertation will analyze data from the SARS-CoV-2 epidemic in California in populations of various sizes. In Chapter 3, we will estimate the effective reproduction number from case and test data for the fifteen most populous counties in California. An example from Orange County is shown in Figure 1.1.

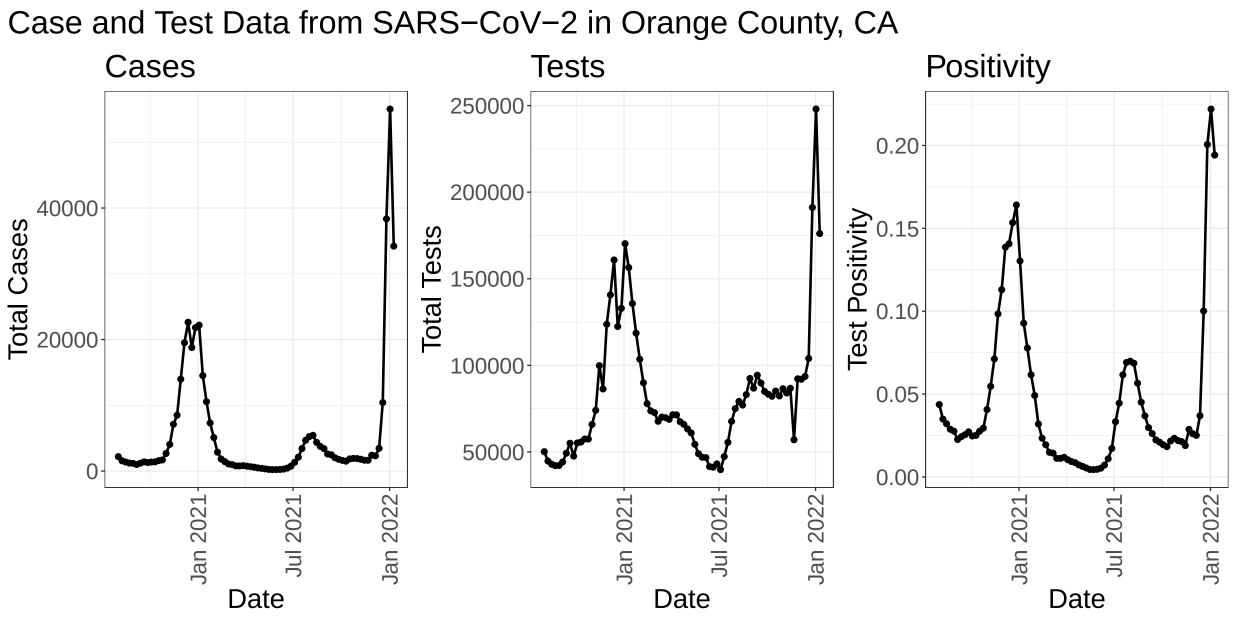


Figure 1.1: Cases (positive tests), total tests (positive and negative tests) and test positivity (cases divide by total tests) for the SARS-CoV-2 epidemic in Orange County, CA from August 2020 through January 2022. Data were collected at a daily time-scale and then aggregated to a weekly level.

We can think of case data as being noisy realizations of the underlying number of newly infectious individuals. The trouble is that reporting rates for cases can change for a variety of reasons unrelated to changes in the number of new infections. For instance in December of 2020, there is a marked decline in the number of new cases near that winter's peak. When we examine the total number of tests administered, however, it is clear that this dip is a result of fewer tests being administered as compared to previous weeks, likely due to the Christmas holiday. When we examine test positivity instead of cases alone, we see an increase in test positivity that week. In Chapter 3, we develop a case-based method for estimating the effective reproduction that incorporates total tests administered as a model covariate based on a branching process inspired model of the epidemic. This work has been published in Goldstein et al. [2024].

1.4.2 SARS-CoV-2 Wastewater Data Collected in Los Angeles County, California

Wastewater data were collected from the Joint Water Pollution Control (JWPCP) wastewater treatment plan in Los Angeles County from July 2021 through February 2022. The JWPCP plant serves 4.8 million residents spread throughout Los Angeles County. It is typical for multiple measurements of the concentration to be taken on a particular day, these multiple measurements are often called replicates. While it is most common to report averages of replicates, we instead will work with the replicates directly. The data are visualized in Figure 1.2.

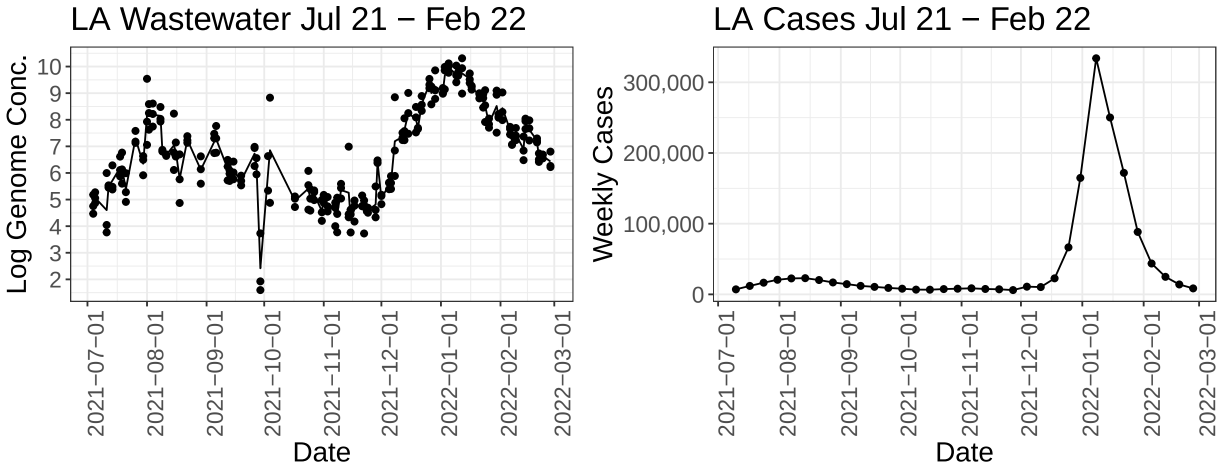


Figure 1.2: Wastewater and case data for the SARS-CoV-2 epidemic in Los Angeles County. Wastewater data were collected approximately every two days, with usually three measurements taken per day, each dot represents a measurement, the line is the average of the measurements. Cases were aggregated to a weekly time scale.

We can see that the wastewater data are quite noisy, there can be outliers and also great variability amongst the replicates on a particular day. On the other hand, there are clear patterns in the data, with the peaks in wastewater concentration aligning well with the peaks in cases. In Chapter 4, we will develop a method to estimate the effective reproduction number from wastewater data by modeling the wastewater data as noisy realizations of the number of currently infectious and recently recovered individuals. We will abandon the branching process model, which does not easily track the number of currently infectious individuals, and instead adapt a deterministic compartmental model that performs well in large populations.

1.4.3 SARS-CoV-2 Wastewater Data Collected at University of California, Irvine

While wastewater data are often collected at large treatment plants such as the JWPCP plant in Los Angeles, it can be collected at many different points of the sewer system, for instance outside university dorms or long-term care facilities [Acer et al., 2022, Keck et al., 2024]. This is potentially very useful, as we expect that epidemic transmission dynamics vary across spatial locations. Wastewater data were collected at University of California, Irvine between January 2022 and June 2022 at a variety of locations across campus corresponding to different residential dormitories housing both graduates and undergraduates. The data were again collected every few days, with usually three replicates available. Case data were also available, reported at lower spatial granularity than the wastewater. These data are displayed in Figure 1.3.

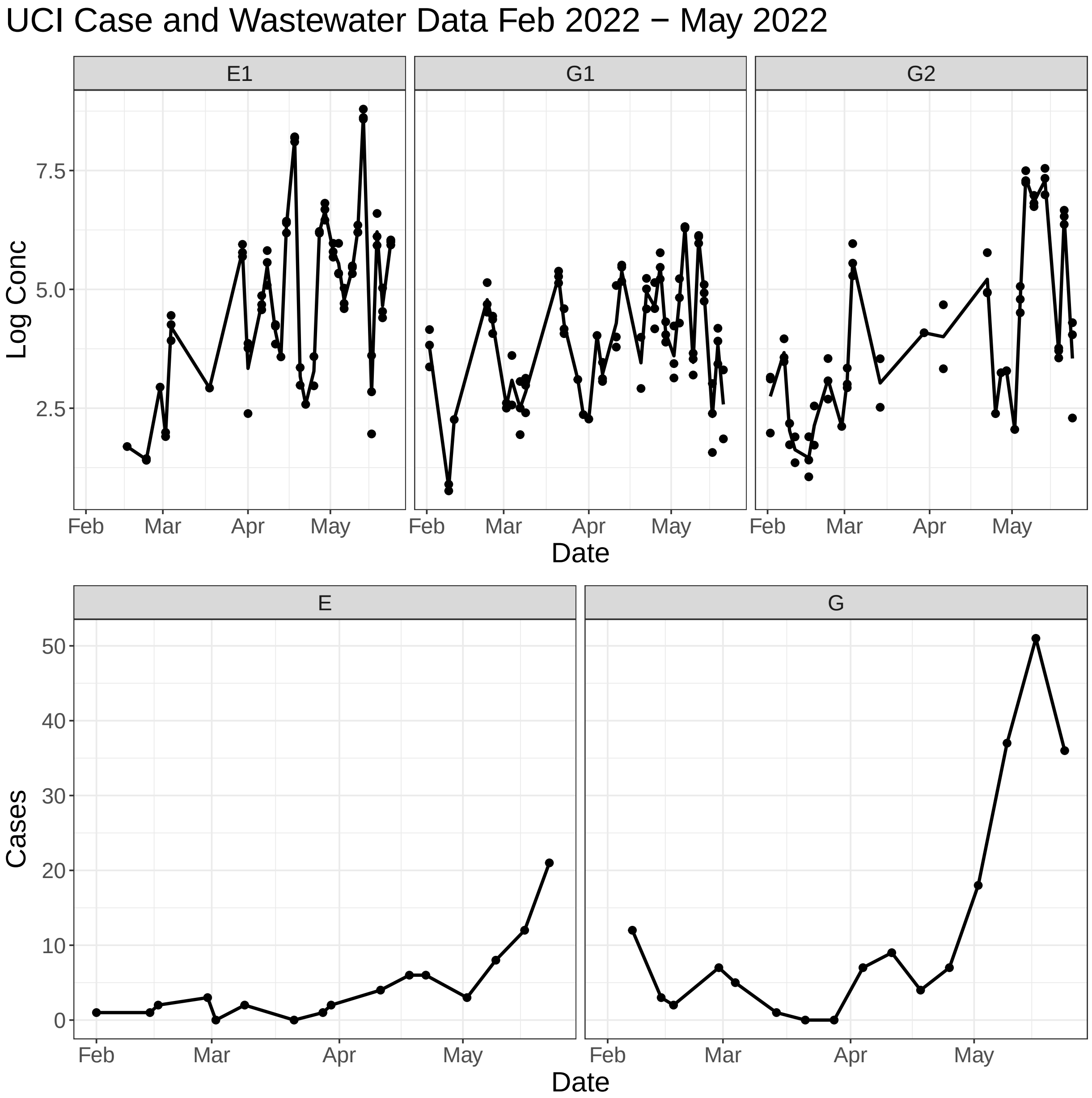


Figure 1.3: Log concentrations of SARS-CoV-2 RNA and weekly reported new COVID-19 cases at UC Irvine for February 2022 through May 2022. For the log concentrations, the dots are individual replicates, and the lines are the mean. The cases are reported at a lower spatial resolution than the concentrations.

G1 and G2 are sub-communities within the larger G community, likewise E1 is a sub-community of E. These sub-communities had around 1000 residents. With such small populations, we are concerned about the suitability of a deterministic model of the epidemic.

In Chapter 5, we develop a model that uses a stochastic model of the epidemic in order to estimate the effective reproduction number from wastewater data.

1.5 Overview of the Dissertation

In Chapter 2, we review concepts in statistics and probability that drive the methodologies developed in later chapters. In particular we review Bayesian inference, Markov jump processes, various epidemic models, including branching and compartmental models, as well as equivalents of the strong law of large numbers and the central limit theorem for Markov jump processes.

In Chapter 3, we develop a branching process inspired method to estimate the effective reproduction number from cases and tests. We develop a new model that incorporates the number of diagnostic tests as a surveillance model covariate. Using simulated data and data from the SARS-CoV-2 pandemic in California, we demonstrate that incorporating tests leads to improved performance over the state-of-the-art.

In Chapter 4, we develop a compartmental model based method to estimate the effective reproduction number from wastewater data. We propose a model where new infections arrive according to a time-varying immigration rate that can be interpreted as an average number of secondary infections produced by one infectious individual per unit time. This model allows us to estimate the effective reproduction number from concentrations of pathogen genomes while avoiding difficult to verify assumptions about the dynamics of the susceptible population. As a byproduct of our primary goal, we also produce a new model for estimating the effective reproduction number from case data using the same framework. We test this modeling framework in an agent-based simulation study with a realistic data generating mechanism that accounts for the time-varying dynamics of pathogen shedding. Finally, we apply our new model to estimating the effective reproduction number of SARS-CoV-2 in Los Angeles, California, using pathogen RNA concentrations collected from a large wastewater treatment facility.

In Chapter 5, we modify the model developed in Chapter 4 for small population settings. In small populations, ideally we would estimate the effective reproduction number using a Markov jump process (MJP) model of the spread of infectious disease, but in practice this is computationally challenging. We propose a simplified MJP that tracks only latent and infectious individuals. Taking advantage of the model's simpler structure, we calculate closed form solutions for the conditional first and second moments, and apply well known analogues of the central limit theorem for MJPs to approximate transition densities as normal, making Bayesian computation tractable. Using simulated pathogen RNA concentrations collected from wastewater data, we demonstrate the advantages of our stochastic model against deterministic counterparts for the purpose of estimating effective reproduction number dynamics. We apply our new model to estimating the effective reproduction number of SARS-CoV-2 in several college campus dormitories at University of California, Irvine.

Chapter 6 concludes the dissertation with a summary and discussion of future directions.

Chapter 2

Background

2.1 Bayesian Inference

There are both Bayesian and frequentist methods for estimating parameters driving infectious disease epidemics, but this disseration will cover only Bayesian methods. Thus, we begin with a brief overview of Bayesian inference.

In Bayesian inference we quantify the range of plausible values of parameters $\mathbf{\Theta}$ given some set of observed data $\mathbf{Y}$ through what is known as the posterior distribution. We treat the parameters $\mathbf{\Theta}$ as random, and require a prior distribution with density $P(\mathbf{\Theta})$ that describes our beliefs about the plausible values of the parameters before having seen the observed data. We also define a likelihood $P(\mathbf{Y}|\mathbf{\Theta})$ that describes how the data $\mathbf{Y}$ are generated given a fixed set of parameters $\mathbf{\Theta}$. Appealing to Bayes' Rule, the density of the posterior distribution is defined as

$$P(\mathbf{\Theta}|\mathbf{Y}) = \frac{P(\mathbf{Y}|\mathbf{\Theta})P(\mathbf{\Theta})}{P(\mathbf{Y})} \tag{2.1}$$

where the denominator $P(\mathbf{Y})$ can be rewritten as an integral $P(\mathbf{Y}) = \int P(\mathbf{Y}|\boldsymbol{\Theta})P(\boldsymbol{\Theta})d\boldsymbol{\Theta}$.

However, this integral is often analytically intractable. In this situation, we can make use of the Markov chain Monte Carlo (MCMC) algorithm to approximate summaries of the posterior distribution (such as the posterior mean or posterior 95% quantile) by constructing a Markov chain whose stationary distribution is the posterior distribution of interest. At each iteration of the algorithm, a proposed value $\boldsymbol{\Theta}'$ is sampled from a proposal distribution q. The proposed value is then accepted with probability

$$p = \min\left\{1, \frac{P(\boldsymbol{\Theta}'|\mathbf{Y})q(\boldsymbol{\Theta}|\boldsymbol{\Theta}')}{P(\boldsymbol{\Theta}|\mathbf{Y})q(\boldsymbol{\Theta}'|\boldsymbol{\Theta})}\right\},$$

which, using Equation 2.1, reduces to

$$p = \min\left\{1, \frac{P(\mathbf{Y}|\boldsymbol{\Theta}')P(\boldsymbol{\Theta}')q(\boldsymbol{\Theta}|\boldsymbol{\Theta}')}{P(\mathbf{Y}|\boldsymbol{\Theta})P(\boldsymbol{\Theta})q(\boldsymbol{\Theta}'|\boldsymbol{\Theta})}\right\}.$$

If the proposed state is rejected, the algorithm remains at its current state (see Chapter 11 in Gelman et al. [2014] for a thorough introduction to MCMC). Note that, to use this algorithm, the likelihood $P(\mathbf{Y}|\boldsymbol{\Theta})$ must be evaluated at each iteration, which will be a major computational obstacle in our application area.

In Chapters 3–5 we will use the MCMC technique known as Hamiltonian Monte Carlo (HMC) that proposes new states of the Markov Chain according to a series of physics-based equations derived from Hamiltonian dynamics [Betancourt, 2017, Neal, 2011]. HMC is known to perform well in high-dimensional settings where other MCMC methods often fail. A key limitation of HMC is that it cannot be used with models with latent discrete random variables, as it relies upon taking the gradient of the proportional posterior, which only exists for continuous latent random variables.

2.2 Markov Jump Processes

Epidemics are often modeled as Markov chains, in particular as continuous time Markov chains. This is particularly useful in statistics, as it allows for data to be collected at irregular intervals. This section provides a brief overview of continuous time Markov chains.

We say that $X(t)$ is a continuous time Markov Chain, or Markov jump process (MJP) if for $t > 0$ and $s \geq r \geq 0$

$$P\left(X(t+s) = j | X(s) = i, X(r) = x(r)\right) = P\left(X(t+s) = j | X(s) = i\right),$$

where $X(t)$ has a state space that is either finite or countably infinite. Intuitively, given the past history of the process, the distribution of the future depends solely on the most recently known state. This is known as the Markovian property. An MJP is homogeneous when $p_{ij}(t) = P(X(t+s) = j | X(s) = i)$ does not depend on s. We define the matrix $\mathbf{P}(t) = \{p_{ij}(t)\}$ to be the transition probability matrix. If the state space of $X(t)$ is of finite size $\mathcal{S}$, then $\mathbf{P}(t)$ is an $\mathcal{S} \times \mathcal{S}$ matrix. By construction, $\mathbf{P}(0) = \mathbf{I}$.

It is often useful to characterize an MJP through its infinitesimal generator matrix. Define $\lambda_i = \lim_{h \to 0^+} \frac{1 - p_{ii}(h)}{h}$. Define $\lambda_{ij} = \lim_{h \to 0^+} \frac{p_{ij}(h)}{h}$. Finally, let $\lambda_{ii} = -\lambda_i$. The infinitesimal generator matrix is $\mathbf{\Lambda} = \{\lambda_{ij}\}$.

An MJP is stable if $\lambda_i < \infty$ and conservative when $\lambda_i = \sum_{j \neq i} \lambda_{ij}$. We will only consider stable and conservative MJPs. In this case, we can write

$$p_{ii}(t + dt) = 1 - \lambda_i dt + o(dt),$$
$$p_{ij}(t + dt) = \lambda_{ij} dt + o(dt).$$

An important consequence of the Markovian property is that the waiting time to the next state in an MJP is exponentially distributed, where the individual waiting time to each type of possible transition is distributed exponentially with rate λ_{ij}.

Generator matrices are often quite sparse, and so for convenience we often describe a MJP in terms of its infinitesimal transition probabilities or infinitesimal rates. For instance, in a pure birth process, at an infinitesimal level there are only two possible transitions, either the population size increases by 1, or it stays the same.

We state without proof the well-known Chapman-Kolmogorov equation that serves as an important building block in describing the properties of MJPs.

$$\mathbf{P}(\mathbf{t}+\mathbf{s}) = \mathbf{P}(t)\mathbf{P}(s). \tag{2.2}$$

Using Chapman-Komogorov, we can show that

$$\mathbf{P}'(\mathbf{t}) = \lim_{h\to 0^+} \frac{\mathbf{P}(t+h)-\mathbf{P}(t)}{h} = \lim_{h\to 0^+} \frac{\mathbf{P}(t)(\mathbf{P}(h)-\mathbf{I})}{h} = \mathbf{P}(t)\boldsymbol{\Lambda}.$$

This is known as the Kolmogorov Forward-Equation. By switching the order of multiplication, we arrive at the Kolmogorov Backward-Equation: $\mathbf{P}'(\mathbf{t}) = \boldsymbol{\Lambda}\mathbf{P}(t)$. In combination with the initial condition $\mathbf{P}(0) = \mathbf{I}$, we can solve for $\mathbf{P}(t)$:

$$\mathbf{P}(t) = \exp(\boldsymbol{\Lambda}) = \sum_{n=0}^{\infty} \frac{\boldsymbol{\Lambda}^n t^n}{n!}.$$

For further reading on Markov jump processes, we refer readers to chapter 8 of Karlin [1966] and chapter 6 of Ross [1985].

2.3 Diffusion Processes

2.3.1 Brownian Motion

We provide a heuristic derivation of Brownian motion based on the limit of a symmetric random walk drawing from Chapter 16 in Feller [1957] and Chapter 10 in Karlin [1966], and leave more rigorous definitions to Chapter 1 of Oksendal [2013] and Chapter 3 of Fuchs [2013]. Consider a series of independent Bernoulli random variables such that

$$P(B_i = 1) = P(B_i = -1) = \frac{1}{2}.$$

Let ΔB be a scalar constant representing the jump size of a random walk and Δt be a time step size. Now define the continuous process

$$W(t) = \Delta B(B_1 + \cdots + B_{\lfloor t/\Delta t \rfloor}).$$

In other words, $W(t)$ is a random walk process where in each time interval Δt, the process moves by length ΔB positively or negatively with equal probability. It makes the jumps up or down by ΔB at times $\Delta t, 2\Delta t, 3\Delta t, \ldots$ and stays constant in intervals $[0, \Delta t), [\Delta t, 2\Delta t), \ldots$

By definition, $E[W(t)] = 0$ and $\mathrm{Var}(W(t)) = (\Delta B)^2 \lfloor \frac{t}{\Delta t} \rfloor$. Now we let $\Delta B \to 0$ and $\Delta t \to 0$ but require that $\frac{\Delta B}{\sqrt{\Delta t}} = 1$. Then the variance of $W(t)$ becomes t.

In this case, the central limit theorem applies, and $W(t)$ converges in distribution to $N(0, t)$. This is standard Brownian motion, it is characterized by the following properties:

1. $W(0) = 0$,

2. $W(t)$ is continuous,

3. $W(t)$ has independent increments,

4. $W(t+s) - W(s) \sim N(0,t)$.

2.3.2 Diffusion Processes

In Chapter 5, we will approximate an MJP with a diffusion process, a kind of continuous valued stochastic process. Thus, we describe the construction of diffusion processes in the following section.

Somewhat oversimplifying the matter, we can think of diffusion processes as solutions to stochastic differential equations. A stochastic differential equation is an extension of a deterministic differential equation, and while the details of the definition are non-trivial, the heuristic definition (taken from Chapter 3 of Oksendal [2013]) is relatively straightforward. We want the solution to a system that looks like

$$\frac{dV}{dt} = \mu(V,t) + \sigma(V,t) \times \text{random noise.}$$

Brownian motion is the best choice for characterizing the noise term, and it is easier to conceptualize this in terms of a discrete version of the above equation:

$$V(t_{k+1}) - V(t_k) = \mu(V(t_k), t_k)(t_{k+1} - t_k) + \sigma(V(t_k), t_k)(W(t_{k+1}) - W(t_k)),$$

where, recall, $W(t_{k+1})$ is standard Brownian motion. If we let $(t_{k+1} - t_k)$ go to 0, we arrive at

$$dV(t) = \mu(V(t), t)dt + \sigma(V(t), t)dW(t).$$

Integrating on both sides yields

$$V(t) = V(0) + \int_0^t \mu(V(s), s)ds + \int_0^t \sigma(V(s), s)dW(s),$$

where the first integral is the usual Lebesgue integral, and the second integral is the stochastic Itô integral. The solution to this equation is our diffusion process, its relevant characteristics are that it is a real valued continuous time Markov process, and if $\sigma(V, t)$ is not a function of V then the transition densities of V will be Gaussian. Mathematically more rigorous introductions are available in Chapter 3 of Oksendal [2013] and Chapter 3 of Fuchs [2013].

2.4 Mathematical Models of the Spread of Infectious Disease

2.4.1 Agent-Based Models

Epidemics are composed of individuals interacting and infecting each other, thus it makes sense to begin with individual level (or agent-based) models of the spread of infectious disease [Andersson and Britton, 2012, Ball and Siri, 2019, Fintzi et al., 2017]. It is common practice to describe the states of an infection, for simplicity we will discuss just three states: susceptible (S), infectious (I), and recovered (R). Many extensions are possible, the most common of which is the inclusion of a latent state (E), that describes when individuals are infected but themselves not yet infectious. We will assume the population is of fixed size N, and further that the population is homogeneously mixing (individuals are equally likely to interact with any other member of the population). Let β/N be the per-capita contact rate, and ν be the recovery rate. When an infected individual makes contact with a susceptible person, the susceptible person becomes infectious. Because the time to recovery is exponentially distributed, the average time spent infectious is $1/\nu$.

We define a Markov chain $\mathbf{X}(t)$ which is a vector of length N, where the entries of $\mathbf{X}(t)$ are the states of each individual in the population. Let I be the total number of infectious individuals at time t. Then the infinitesimal transition rates between states $\mathbf{X}$ and $\mathbf{X}'$ are

$$\lambda_{\mathbf{X}\mathbf{X}'} = \begin{cases} \frac{\beta}{N}I, & \text{if, for a single subject } j \ X_j = S \text{ and } X'_j = I \\ \nu, & \text{if, for a single subject } j \ X_j = I \text{ and } X'_j = R. \end{cases}$$

2.4.2 Compartmental Models

This dissertation is concerned with the (far more common) situation where individual level data is unavailable, in which case models of epidemics spreading through populations, rather than amongst individuals, are more germane.

Happily, we can explicitly connect the individual-level SIR model, and the population level SIR model, often called a compartmental model. The population level SIR model counts the number of individuals in each state at time t, so it is a three dimensional MJP: $\mathbf{G}(t) = \{S(t), I(t), R(t)\}$. We can derive the rates for the population level model by recalling that the minimum of a set of independent exponential random variables is itself an exponential random variable with rate equal to the sum of the rates. More technical discussion connecting the two versions of the models can be found in Fintzi [2018]. In the population level SIR, we are interested in the next time any member of the population changes states, in other words, the minimum of all of the possible changes in states. At time t, if we have S total susceptible individuals and I total infectious individuals, then summing up the individual rates yields

$$\lambda_{\mathbf{G}\mathbf{G}'} = \begin{cases} \frac{\beta}{N} IS, \quad \text{If } \mathbf{G}' = \{S(t)-1, I(t)+1, R(t)\}, \\ \nu I \quad \text{If } \mathbf{G}' = \{S(t), I(t)-1, R(t)+1\}. \end{cases}$$

Figure 2.1 is a diagram of the population level SIR model.

$$\begin{tikzcd}
 & (S-1, I+1, R) \\
(S, I, R) \arrow[ur, "(\beta/N)IS"] \arrow[dr, "\nu I"'] & \\
 & (S, I-1, R+1)
\end{tikzcd}$$

Figure 2.1: Diagram of the population-level SIR model. A population is divided into three states, Susceptible (S), Infected (I) and Recovered (R). On an individual level, individuals can move from S to I, and from I to R, but no other moves between states are possible. The population-level model tracks the total number of individuals in each of the three states at any given time. The states are often called compartments.

Compartmental models are the most commonly used epidemic models. There are many introductions to the topic, including Allen [2010], Blackwood and Childs [2018], Keeling and Rohani [2008], Andersson and Britton [2012] and Renshaw [2015] among others. Using the rate parameters above, we can write the infinitesimal probabilities of the SIR model as

$$P(\mathbf{G}(t+dt) = (S-1, I+1, R) \mid \mathbf{G}(t) = (S, I, R)) = \beta I(S/N)dt + o(dt),$$
$$P(\mathbf{G}(t+dt) = (S, I-1, R+1) \mid \mathbf{G}(t) = (S, I, R)) = \nu I dt + o(dt),$$
$$P(\mathbf{G}(t+dt) = (S, I, R) \mid \mathbf{G}(t) = (S, I, R)) = 1 - (\beta I(S/N) + \nu I)dt + o(dt).$$

Because the population size is fixed at N, the model can be fully expressed by any two of the compartments, and thus the cardinality of the state-space of the model $|\mathcal{S}|$ is of the order N^2. Calculating the matrix exponential of $\mathbf{\Lambda}$ becomes computationally intensive at even relatively small population sizes, making this system difficult to work with in practice [Rupp et al., 2024, Ho et al., 2018, Moler and Van Loan, 2003].

Often, practitioners use the deterministic version of the SIR model, where changes in the system are described by a series of ordinary differential equations

$$\begin{aligned}\frac{dS}{dt} &= -\beta \times I \times S/N,\\ \frac{dI}{dt} &= \beta \times I \times S/N - \nu \times I,\\ \frac{dR}{dt} &= \nu \times I.\end{aligned}$$

Note here that the compartments are now continuous, but $S(t) + I(t) + R(t) = N$ for any t, preserving the closed population. This system is non-linear, and does not have a closed form solution, but can be numerically solved using common ODE numerical solvers. The deterministic ODE system can be thought of as the limit of the MJP as the population size (N) diverges to infinity, which we discuss in greater detail in Section 2.4.5.

Two important quantities in infectious disease models are the basic reproduction number and the effective reproduction number. The basic reproduction number, R_0, is the average number of individuals an infectious person would infect in a completely susceptible population. For the SIR model, $R_0 = \beta/\nu$.

We provide some intuition about R_0 for the deterministic SIR model based on Blackwood and Childs [2018]. Consider the situation with one infected individual at time 0, with all other individuals in the susceptible population. Then

$$\frac{dI}{dt}|_{t=0} = \beta/N \times (N-1) \times 1 - \nu \times 1.$$

We are interested in the situation when the derivative is greater than 0 or less than 0, i.e. when the size of the infectious compartment will increase or decrease. We can write the inequality $\frac{dI}{dt}|_{t=0} < 0$ as

$$\beta/\nu \times (N-1)/N < 1.$$

Suppose $(N-1)/N$ is large enough to be close to 1, then we have

$$\beta/\nu < 1,$$

as the quantity of interest. This is the quantity R_0, and when R_0 is larger than 1, we expect to have an outbreak, when it is below 1 we do not expect to have an outbreak. For more general deterministic models it can be defined using the next-generation matrix [Diekmann et al., 2010].

In the stochastic case, for a homogeneously mixing population, R_0 is defined using a branching process approximation where infectious individuals create new infectious individuals in an independent and identically distributed fashion, which can be a good approximation in the early stages of an outbreak when the number of infections is small and the number of susceptibles is large [Andersson and Britton, 2012, Ball and Donnelly, 1995]. In the branching process, R_0 is the expected number of infections created by one infectious person, and governs the probability of an outbreak occurring.

The effective reproduction number $R_t = R_0 \times S(t)/N$ is the average number of individuals an infectious person infected at time t would subsequently infect assuming that the susceptible proportion remained the same as at time t. The effective reproduction number is simply a modified basic reproduction number that corrects for the fact that the susceptible population has been depleted. This makes it a useful quantity for describing the current conditions of the outbreak.

To highlight the difference between the stochastic and deterministic systems, consider a setting where we begin the epidemic with one infectious individual and R_0 is larger than 1. This means that the average number of individuals the infectious person would infect in a completely susceptible population is larger than 1. If we simulate a deterministic SIR with $R_0 > 1$ for long enough, then the number of infectious individuals will be greater than 1. However, in a stochastic SIR model, it is possible for us to simulate an outbreak where no other individuals are infected, simply by chance. In general, we think of the stochastic SIR as being closer to how epidemic outbreaks actually evolve, but the deterministic system as being a good enough approximation for large populations.

2.4.3 Motivation for Approximation

Suppose that each week a perfectly administered survey is sent to A people to find out who in the population is currently infectious with some disease (a more realistic version of this type of survey was administered in the United Kingdom during the SARS-CoV-2 pandemic [Pouwels et al., 2021]). Suppose further that we think the disease in question is well modeled with an SIR model $\mathbf{G}(t)$. Let $\mathbf{G}_i = \mathbf{G}(t_i)$, and likewise $I_i = I(t_i)$. We might model the data Y_i, collected at time t_i, as a Binomial random variable, where

$$Y_i | I_i \sim \text{Binomial}(A, I_i/N).$$

That is, the probability that someone tested positive on the survey is equal to the proportion of the population that is currently infectious. Conditional on the states $\mathbf{G}_1, \ldots \mathbf{G}_M$, we will assume the Ys are independent of each other. Suppose we have M such surveys. We are interested in the posterior distribution $P(\beta, \nu | \mathbf{Y})$. Then,

$$P(\beta, \nu | \mathbf{Y}) \propto P(\mathbf{Y} | \beta, \nu) P(\beta, \nu).$$

Unfortunately, the data are quite dependent on each other, so $P(\mathbf{Y}|\beta,\nu)$ is not a simple product of binomial likelihoods. Figure 2.2 visualizes the dependence structure of the model.

$$\begin{tikzcd}
(S_1, I_1, R_1) \arrow[r] \arrow[d] & (S_2, I_2, R_2) \arrow[r] \arrow[d] & (S_3, I_3, R_3) \arrow[d] \\
Y_1 & Y_2 & Y_3
\end{tikzcd}$$

Figure 2.2: Dependence structure of our hypothetical SIR model using Binomial prevalence data. Conditioned on the state of the SIR model at each time point, the observed data Y are independent of each other. The current state of the SIR model depends on the state of the SIR model at the previous time point.

Instead

$$P(\mathbf{Y}|\beta,\nu) = \sum_{\mathbf{G}_M} \cdots \sum_{\mathbf{G}_0} P(\mathbf{Y}|\beta,\nu,\mathbf{G}_0,\ldots\mathbf{G}_M)P(\mathbf{G}_0,\ldots,\mathbf{G}_M|\beta,\nu).$$

When the state-space of $\mathbf{G}$ is large, this sum is computationally intractable, and in the case of the SIR model, the state space is $\mathcal{O}(N^2)$. For any reasonably sized human population, this integral will be intractable. Taking advantage of the Markovian property and our assumed data model, we could instead re-write the integral as

$$P(\mathbf{Y}|\beta,\nu) = \sum_{\mathbf{G}_M} \cdots \sum_{\mathbf{G}_0} \prod_{i=1}^{M} P(Y_i|\mathbf{G}_i)P(\mathbf{G}_i|\mathbf{G}_{i-1})P(\mathbf{G}_0).$$

In such a way, we could approximate this integral through data augmentation in an MCMC algorithm, where we adjust our posterior of interest to be the posterior $P(\beta,\nu,\mathbf{G}_0,\ldots\mathbf{G}_M|\mathbf{Y})$. However, to do this, we would need to be able to evaluate the transition probabilities for $P(\mathbf{G}_i|\mathbf{G}_{i-1})$ in the accept/reject step of our MCMC algorithm. These transition probabilities do not exist in closed form for the SIR model, and while progress has been made on computational solutions [Ho et al., 2018, Rupp et al., 2024], the computations still take, at best,

$\mathcal{O}(N^2)$ time, and the computational costs remain formidable. There are a few approaches to address this problem, but the one we will take in this dissertation is to choose an approximate process such that the transition probabilities $P(\mathbf{G}_i|\mathbf{G}_{i-1})$ become once again tractable. What follows is a discussion of the approximate processes we will use in this dissertation.

2.4.4 Branching Process Models

One of the simplest approximate models we could use is a branching process model, where we assume that infectious individuals create a random total number of new infections in an independent and identically distributed (IID) manner unrestricted by a finite supply of susceptible people to infect. As mentioned previously, this is a reasonable approximation early in an epidemic [Andersson and Britton, 2012, Ball and Donnelly, 1995]. One can then leverage properties of branching processes to derive useful equations describing the expected total number of individuals who have been infected through this branching process, called cumulative incidence, in terms of the past cumulative incidence. From the cumulative incidence, one can derive an equation describing changes in the expected number of new infections, or incidence, which naturally can be tied to common data sources such as the observed number of new cases. This equation is often called the renewal equation. Define J_i to be the new cases generated in $(t_{i-1}, t_i]$, g_u to be the discretized generation time pdf (in the branching process this is the pdf of the distribution of the time between the birth of the infectious individual and the time when it creates all of its offspring), and R_t the effective reproduction number. The renewal equation is

$$J_t = R_t \sum_{u=0}^{t-1} J_{t-u} g_u.$$

Intuitively, the renewal equation posits that the number of new infections is the product of the effective reproduction number and a weighted sum of the previous new cases, where the weights are such that those at the peak of their infectiousness will contribute more to new incidence.

In point of fact, the development of these methods in the real world is much the opposite of what we have described. Fraser popularized the idea of using equations relating incidence to past incidence without referencing any branching process connections formally [Fraser, 2007], which gained further popularity when the idea was incorporated into the widely used `EpiEstim` package for estimating R_t [Cori et al., 2013]. More recently, Pakkanen et al. [2023] provided a formal justification for this approach based on Crump-Mode-Jaegers processes [Crump and Mode, 1968, 1969, Jagers, 1975], which we will summarise.

Let the first individual in the outbreak be infected at time $\tau \geq 0$, and behave according to the triple $\{L^\tau, \chi^\tau, C^\tau\}$. The variable L^τ is a positive random variable describing how long the individual is infectious with CDF G and pdf g. The variable $\chi^\tau(u)$ is an indicator variable that equals 1 if the individual has been infected by time τ. The variable C^τ is a counting process that describes how many other infected individuals are generated by the infectious individual. For simplicity, consider the case when

$$C^\tau(u) = \begin{cases} 0, \quad u < L^\tau, \\ \xi(\tau + L^\tau), \quad u \geq L^\tau, \end{cases}$$

where $\xi(.)$ is some stochastic process taking integer values that is independent of L^τ. If $\xi(.)$ does not depend on time, this is then a classic Bellman-Harris process, whereas when it is time dependent it is the time-varying Bellman-Harris process [Kimmel, 1983]. In this model, all individuals infected by the first individual are infected simultaneously at time L^τ, and L^τ is the generation time, the time between when the first individual was infected, and the time they infected (generated) the next generation of infected individuals. The expectation of ξ, $E[\xi(t)] = R(t)$ is defined as the reproduction number, and is the branching process

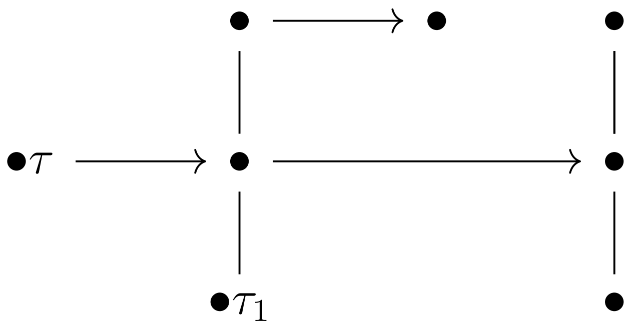


Figure 2.3: Visualization of a Bellman-Harris process. The first individual is infected at time τ. They then produce all of their offspring (new infections) at time τ_1. The times at which offspring produce have a distribution, as does the number of offspring an individual produces. If the distribution for the number of offspring is time-varying, then the average number of offspring is the effective reproduction number. A key property of the branching process is that, because the individuals are assumed to be IID, the process beginning at τ_1 has the same distribution as the process beginning at τ.

equivalent of the effective reproduction number. Figure 2.3 visualizes the Bellman-Harris process.

Let $\mathcal{J}$ be the set of infected individuals, then we are interested in the process

$$Z(t,\tau) = \sum_{i \in \mathcal{J}} \chi_i^{\tau_i}(t-\tau_i), t \geq \tau \geq 0,$$

where τ denotes the time the first case was infected. Then $Z(t,\tau)$ describes the cumulative number of individuals who have ever been infected during the epidemic, called the cumulative incidence. Let $f(t,\tau) = E[Z(t,\tau)]$, $\mathcal{J}^*$ index all infected individuals except for the first individual, and k be the number of individuals infected by the first infected individual up to time t. The key insight is that we can index $\sum_{i \in \mathcal{J}^*} \chi_i^{\tau_i}(t-\tau_i)$ by the k offspring of the original infected individual infected by time t so that

$$f(t,\tau) = E\left[\chi^{\tau}(t-\tau)\right] + E\left[\sum_{k:\tau_{i_k} \leq t} \sum_{i \in \mathcal{J}_k^*} \chi_i^{\tau_i}(t-\tau_i)\right].$$

But because the individuals are all independent and identically distributed, we can think of each of these sums as a new branching process, exactly the same as the original but starting at the infection time τ_{i_k}. Then, we can write

$$f(t,\tau) = E[\chi^\tau(t-\tau)] + E\left[\sum_{k:\tau_{i_k}\leq t} f(t,\tau_{i_k})\right].$$

We now re-index the sum in terms of ν and $\Delta C^\tau(u) = C^\tau(u) - \lim_{\nu\to u^-} C^\tau(\nu)$ which is the number of infected individuals created by the first infected individual at time u. If the τ_is are different, this will always be a one, if the τ_i occur simultaneously, as in the Bellman-Harris process, this will be equal to k. It is clear that

$$\begin{aligned} E\left[\sum_{k:\tau_{i_k}\leq t} f(t,\tau_{i_k})\right] &= E\left[\sum_{\nu\in(\tau,t]} f(t,\nu)\Delta C^\tau(\nu-\tau)\right] \\ &= E\left[\sum_{u\in(0,t-\tau]} f(t,u+\tau)\Delta C^\tau(u)\right] \\ &= E\left[\int_0^{t-\tau} f(t,u+\tau)dC^\tau(u)\right]. \end{aligned}$$

Swapping the order of integration and expectations (allowable assuming some regularity conditions) brings us to

$$f(t,\tau) = E[\chi^\tau(t-\tau)] + \int_0^{t-\tau} f(t,u+\tau)dE[C^\tau(u)].$$

Returning to our Bellman-Harris process, defining $I\{\}$ to be the indicator function, and recalling that C^τ and L^τ are independent, we have

$$\begin{aligned} E[C^\tau](u) &= E[\xi(L^\tau+\tau)I\{L^\tau\leq u\}] \\ &= E\left[E[\xi(L^\tau+\tau)|L^\tau]I\{L^\tau\leq u\}\right] \\ &= \int_0^u R(v+\tau)g(v)du. \end{aligned}$$

Then, the final result (for $t \geq \tau$) is

$$f(t,\tau) = 1 + \int_0^{t-\tau} f(t, u+\tau) R(u+\tau) g(u) du.$$

We define the expected incidence $J(t,\tau)$ as

$$J(t,\tau) = \frac{df(t,\tau)}{dt}.$$

By taking advantage of the recursive relationship we just defined, as well as the use of Grönwall's inequality to justify pushing $R(t)$ outside of the integral, Pakkanen et al. [2023] show that

$$J(t,\tau) = \delta(t-\tau) + R(t) \int_0^{t-\tau} J(t-u,\tau) g(u) du,$$

which is essentially the continuous version of the Renewal Equation introduced by Fraser [Fraser, 2007]. Interestingly, χ can be modified to create renewal equations for prevalence instead of cumulative incidence. The counting process C^τ can also be made more complicated to allow for infections to not occur all at once, although this makes defining the effective reproduction number much less trivial [Pakkanen et al., 2023].

2.4.5 Deterministic and Diffusion Limits of MJPs

The branching process approach is appealing because it is simple (although that simplicity can be deceptive, see the discussion by Svensson [2007] on generation times). Another reasonable approach is to adopt an approximation that preserves some of the structure of the original MJP. It is common practice to instead use a deterministic or diffusion version of an MJP, for example we can approximate the stochastic SIR model with the deterministic SIR Model. This turns out to be justified under what amounts to a strong law of large numbers for MJPs, where the value n that diverges to ∞ can be interpreted either as the total size of the MJP when the population is closed (as in the simple SIR model), or as the "typical initial size of the system" if the population has no finite limit [Barbour, 1974]. There is also a corresponding central limit theorem showing that the transition probabilities of MJPs can be approximated with the normal distribution. In practice, if the counts in the compartments being modeled are large enough, a deterministic or diffusion model is an appropriate approximate model.

Rigorous proofs of these theorems go back to Kurtz [Kurtz, 1970, 1971], with Barbour providing some useful clarifications in the case when the population is not necessarily finite [Barbour, 1974]. More legible proofs can be found in Section 2 of Britton and Pardoux [2019], while Chapter 5 of Andersson and Britton [2012] provides a less rigorous but still insightful explanation. Another useful way to build intuition is by following the construction of the linear noise approximation (LNA), which begins by showing how an MJP can be approximated by a stochastic differential equation, and then further linearizing the stochastic differential equation so that the process becomes Gaussian [Wallace et al., 2012]. We will provide informal derivations of both results based on Andersson and Britton [2012].

ODE Approximations to MJPs

It is most useful to begin with a result from Poisson processes. Let $X(t)$ be a Poisson process with rate $\lambda = 1$. Then $\lim_{n\to\infty} \sup_{s\leq t}|n^{-1}X(ns) - s|$ converges to 0 almost surely. Note that

$$n^{-1}X(ns) = n^{-1}\sum_{i=1}^{n}\left(X(t_i) - X(t_{i-1})\right),$$

where for all i, $t_i - t_{i-1} = s$. Since Poisson processes have stationary and independent increments, $X(t_i) - X(t_{i-1}) \sim \text{Poisson}(s)$ and are independent across i. The result follows from the strong law of large numbers.

We will use the same result in the case when s is itself a stochastic process, although proving it is nontrivial and makes heavy use of the fact that $X(t) - t$ is itself a martingale; interested readers should consult Britton and Pardoux [2019].

Define a sequence of MJPs $\mathbf{U_n} = \{\mathbf{U_n}(t); t \geq 0\}$, with a finite number of potential infinitesimal jumps $\mathbf{l_i}$ and infinitesimal rates $nh(n^{-1}\mathbf{U_n})$ so that

$$\begin{aligned} P(\mathbf{U_n}(t+dt) = u + \mathbf{l_i} | \mathbf{U_n}(t)) &= nh_i\left(n^{-1}\mathbf{U_n}\right) dt + o(dt), \\ P(\mathbf{U_n}(t+dt) = u | \mathbf{U_n}(t)) &= 1 - \sum_i nh_i\left(n^{-1}\mathbf{U_n}\right) dt + o(dt). \end{aligned}$$

Set $\mathbf{U_n}(0)$ to be non-random and suppose that as $n \to \infty$, $\mathbf{U_n}(0) \to \mathbf{u_0}$. Also assume h is continuous and well behaved. We can rewrite $\mathbf{U_n}$ in terms of independent Poisson processes $(X_i(t))$ with rate $\lambda = 1$ by

$$\mathbf{U_n}(t) = \mathbf{U_n}(0) + \sum_i \mathbf{l_i} X_i\left(n\int_0^t h_i\left(n^{-1}\mathbf{U_n}(s)\right) ds\right).$$

One way to see the equivalence is to note that the probability of a jump $\mathbf{l}$ in the interval $(t, t+dt]$ (assuming that no other jump occurs so that $\mathbf{U_n}(s) = \mathbf{U_n}(t)$ for the duration of the interval) is $nh\left(n^{-1}\mathbf{U_n}\right) dt$, i.e. the length of the corresponding interval in the Poisson process, which is the same as in the original construction of $\mathbf{U_n}$.

From there we define $\hat{X}_i(t) = X_i(t) - t$, $\bar{\mathbf{U}}_\mathbf{n}(t) = n^{-1}\mathbf{U_n}(t)$, and $F(u) = \sum_i \mathbf{l_i} h_i(u)$. Dividing our equation by n we now have

$$\bar{\mathbf{U}}_\mathbf{n}(t) = \bar{\mathbf{U}}_\mathbf{n}(0) + n^{-1}\sum_i \mathbf{l_i}\hat{X}_i\left(n\int_0^t h_i\left(\bar{\mathbf{U}}_\mathbf{n}(s) ds\right)\right) + \int_0^t F\left(\bar{\mathbf{U}}_\mathbf{n}(s)\right) ds.$$

Recalling our result on Poisson processes, the sum in the middle will tend to zero almost surely, and implies that $\mathbf{\bar{U}_n}(t)$ will converge to some function $\mathbf{u(t)}$ defined by

$$\mathbf{u(t)} = \mathbf{u_0} + \int_0^t F\left(\mathbf{u}(s)\right) ds,$$

or, in differential equation form

$$\frac{d\mathbf{u(t)}}{dt} = F(\mathbf{u(t)}).$$

For the SIR model, n is the population size N, $\mathbf{U_n}(t) = (S(t), I(t))$ (by assumption $R(t) = N - S(t) - I(t)$), $h\left(\mathbf{U_n}(t)\right) = (\beta/NS(t)I(t), \nu I(t))$. Then $n^{-1}\mathbf{U_n}(t) = (S(t)/N, I(t)/N)$ is a process where the total population size is 1, and the rates for $n^{-1}\mathbf{U_n}(t)$ are $h\left(n^{-1}\mathbf{U_n}(t)\right) = (\beta S(t)I(t)/N^2, \gamma I(t)/N)$. It is clear that $h\left(\mathbf{U_n}(t)\right) = nh\left(n^1\mathbf{U_n}(t)\right)$. For the SIR model, $l = ((-1,1),(0,-1))$. Then the deterministic system is

$$\begin{aligned}\frac{u_1(t)}{dt} &= -\beta u_1(t)u_2(t),\\ \frac{u_2(t)}{dt} &= \beta u_1(t)u_2(t) - \nu u_2(t).\end{aligned}$$

Multiplying u by N returns us to the deterministic SIR system.

Stochastic Approximations

We state another property of Poisson processes without proof: $\sqrt{n}\left(n^{-1}X(nt) - t\right) = \sqrt{n}\hat{X}(nt)$ converges in distribution to standard Brownian motion $W(t)$ as $n \to \infty$. This result will also extend to the case when t is stochastic. Applying our techniques from the previous section, we can write

$$\mathbf{V_n}(t) = \sqrt{n}(\mathbf{\bar{U}_n}(t) - \mathbf{u(t)}) = \mathbf{V_n}(0) + \sum_i \mathbf{l_i}\sqrt{n}\hat{X}_i\left(n\int_0^t h_i\left(\mathbf{\bar{U}_n}(s)\right) ds\right) + \int_0^t \sqrt{n}(F\left(\mathbf{\bar{U}_n}(s)\right) - F(\mathbf{u}(s)))ds.$$

We can Taylor expand $F(\bar{\mathbf{U}}_{\mathbf{n}}(s))$ around $\mathbf{u}(s)$ to write

$$\begin{aligned}\sqrt{n}(F(\bar{\mathbf{U}}_{\mathbf{n}}(s)) - F(\mathbf{u}(s))) &= \sqrt{n}\partial F(\mathbf{u}(s))(\bar{\mathbf{U}}_{\mathbf{n}}(s) - F(\mathbf{u}(s))) + O(\sqrt{n}|F(\bar{\mathbf{U}}_{\mathbf{n}}(s)) - F(\mathbf{u}(s))|^2) \\ &= \sqrt{n}\partial F(\mathbf{u}(s))\mathbf{V}_{\mathbf{n}}(s) + O(|F(\bar{\mathbf{U}}_{\mathbf{n}}(s)) - F(\mathbf{u}(s))|)\mathbf{V}_{\mathbf{n}}(s),\end{aligned}$$

where ∂F is the Jacobian of F. We know that $\mathbf{U}_{\mathbf{n}}(t)$ converges to $\mathbf{u}(\mathbf{t})$ almost surely, what we have shown so far then implies that $\mathbf{V}_{\mathbf{n}}(t)$ will converge to $\mathbf{V}(t)$ where

$$\mathbf{V}(t) = \mathbf{v_0} + \sum_i \mathbf{l_i} W_i \left(\int_0^t h_i(\mathbf{u}(s))ds \right) + \int_0^t \partial F\left(\mathbf{u}(s)\right)\mathbf{V}(s)ds, \tag{2.3}$$

which is indeed the case.

Recall that W_i is standard Brownian motion. Of critical importance is that the Brownian motion will not depend on the diffusion process $\mathbf{V}(t)$, but instead only on the deterministic process $\mathbf{u}(t)$. This ensures that the diffusion process $\mathbf{V}(t)$ has Gaussian transition densities.

Deriving the moments of these Gaussian transition densities is a non-trivial task, one approach is to use the LNA [Wallace et al., 2012]. Another approach is to instead approximate (or derive exactly) the moments of the MJP, and use those moments as the moments of the Gaussian density, as in the work of Isham [1991] and Buckingham-Jeffery et al. [2018].

2.5 Other Strategies

We have covered in some depth the approximation strategies used in Chapters 3–5, but there are many other approaches to tackle the problem of inference for epidemic models.

In the time series SIR (TSIR), the infectious period is fixed at a single unit of time, the incidence thus becomes the prevalence, and the process is now in discrete rather than continuous time [Finkenstädt and Grenfell, 2000, Wakefield et al., 2020]. This allows for modeling incidence as a pure birth process with negative binomial transition probabilities.

Another approach is to instead use a linear birth-death process for the number of infectious individuals, so that the infinitesimal rate of a new infectious individual is proportional only to the number of currently infectious individuals, without regard to the number of susceptibles. Cauchemez and Ferguson use the diffusion approximation, a Cox-Ingersoll-Ross process with tractable transition densities [Cauchemez and Ferguson, 2008], and Stadler et. al used this birth-death process to link models of viral evolution to transmission dynamics [Stadler et al., 2013].

Other approaches attempt to make the MJP itself tractable. Returning to the agent-based model, augmenting the posterior with the individual infection and recovery times creates a tractable likelihood [Gibson and Renshaw, 1998, O'Neill and Roberts, 1999, Fintzi et al., 2017]. A nice introduction is available in the lecture notes in Kypraios and O'Neill [2021].

Approximate Bayesian computation can instead be employed, where model parameters are simulated from the prior, MJP trajectories are simulated, and finally observed data are simulated from the MJP. Then, the parameters are accepted as samples from the approximate posterior if the simulated data matches the observed data according to some metric based on summary statistics [Neal, 2020].

The final approach we will mention is particle Markov chain Monte Carlo. This algorithm relies on using sequential Monte Carlo to approximate the observed data likelihood by sequentially importance sampling values for the unobserved MJP trajectory [Andrieu et al., 2010, King et al., 2016]. The technique is likewise quite computationally intensive, as the sequential Monte Carlo algorithm must be run for each MCMC iteration. The lecture notes in King et al. provide an excellent introduction to using sequential Monte Carlo for epidemic models.

Chapter 3

Incorporating Testing Volume into Estimation of Effective Reproduction Number Dynamics

3.1 Introduction

In an infectious disease epidemic, the effective reproduction number is the average number of people a newly infected person will subsequently infect. When the effective reproduction number is above one, an epidemic is out of control and will continue to grow, vice versa if it is below one. This makes the effective reproduction number a useful summary of the state of an epidemic which can provide guidance to policy makers. As such, estimates of the effective reproduction number based on observed data can be an important part of any public health response during an epidemic. Recent examples from the SARS-CoV-2 pandemic include work by Mishra et al. [2020] in Scotland, as well as efforts by Swiss National Covid-19 Science Task Force [2020].

An early effort of using a likelihood based approach to estimate the effective reproduction number is that of Wallinga and Teunis [2004], which is based on modeling transmission trees. A recently popular class of estimators for the effective reproduction number (used in both [Mishra et al., 2020] and [Swiss National Covid-19 Science Task Force, 2020]) is inspired by stochastic branching process models, where infectious individuals infect a random number of new individuals at random points in time. The most widely used model in this class is available in the `EpiEstim` R package [Cori et al., 2013, Thompson et al., 2019], which is based on ideas put forth by Fraser [2007]. `EpiEstim` assumes all new infections (incidence) are observed, and uses a time series of observed cases as data. During the SARS-CoV-2 pandemic, a number of methods in this class of estimators have been developed. The methods of Parag [2021] and Capistrán et al. [2022] continue to assume incidence (or incidence up to a constant) are observed, and focus on improving how changes in R_t are modeled over time, while avoiding Markov chain Monte Carlo based methodologies. The methods of Abbott et al. [2020], Huisman et al. [2022a], Scott et al. [2021], and Bhatt et al. [2023] use more computationally intensive approaches that model observed data as functions of latent incidence, either through explicit Bayesian models [Abbott et al., 2020, Scott et al., 2021, Bhatt et al., 2023] or through a pipeline that first bootstraps latent incidence which is then used as input into `EpiEstim`. The methods of Teh et al. [2022], Scott et al. [2021], and Bhatt et al. [2023] also begin to tackle the problem of how to estimate R_t across spatial locations. Many of these methods have not been scrutinized via extensive simulation studies under model mis-specification, or in some cases, not probed at all, making it difficult to understand the strengths and weaknesses of this class of methods.

This gap in knowledge is particularly relevant when it comes to applying such methods to observed case counts of an infectious disease. As the SARS-CoV-2 pandemic has demonstrated, observed cases of an infectious disease are often circuitously related to the true number of new infections, due to constraints in testing supply, asymptomatic infections, testing eligibility, and reporting delays. These factors can make estimating the effective reproduction number from cases quite difficult in real world situations. This is a widely recognized challenge; a recent survey of papers using EpiEstim found the most common challenge for users was dealing with the quality of observed case data [Nash et al., 2022]. One sensible approach to resolving this issue is to use other sources of data. For instance, Flaxman et al. [2020] used a model similar to those available in epidemia to assess the effects of non-pharmaceutical interventions by fitting a model to death counts rather than case counts, while Mishra et al. [2020] incorporated data sources such as deaths and sero-prevalence data in addition to case data. Turning to other data sources is an appealing strategy for retrospective analyses, but during an ongoing epidemic it is often desirable to provide real time estimates of the effective reproduction number, a task called now-casting. When now-casting, case data is one of the earliest available data sources to indicate a change in the effective reproduction number. It behooves us, then, to develop reasonable methods for using case data when estimating the effective reproduction number, despite the difficulties involved.

Our study has two main contributions. First, we develop our own model for estimating the effective reproduction number making different modeling choices than other available methods. The most significant of these is that we incorporate the number of diagnostic tests administered (both positive and negative) as a covariate in our model. Second, to increase understanding of the broader class of branching process inspired methods, we conduct simulation studies comparing our new model to EpiEstim and a model constructed using the epidemia package developed by Scott et al. [2021]. The latter approach builds on the EpiEstim framework by allowing for more flexible and complex models that treat new infections as unobserved variables, with various time series such as cases or deaths modeled as noisy realizations of unobserved infections used as data [Scott et al., 2021, Bhatt et al., 2023]. In particular, we explore scenarios with differing diagnostic test availability. We also fit our model to real data from the SARS-CoV-2 pandemic in fifteen California counties. Our results show that our new model outperforms existing methodologies under a variety of different testing scenarios and provides novel insights when applied to real data, highlighting the utility of incorporating tests when using case data as well as distributional choices made in the modeling process.

3.2 Methods

3.2.1 Available Data

Consider an outbreak observed for a total of T time intervals. We restrict ourselves to two kinds of infectious disease outbreak data. The first is the time series of observed cases, $\mathbf{O} = (O_1, O_2, O_3, \ldots, O_T)$, where O_u is the number of newly observed cases of an infectious disease during time interval u. The second is the time series of diagnostic tests, $\mathbf{M} = (M_1, M_2, \ldots, M_T)$, where M_u is the total number of diagnostic tests administered during time interval u. For this chapter, we assume tests are perfectly accurate. We also do not model the total number of tests performed, but rather model the number of positive tests conditioned on the total number of tests. We assume that O_u is a noisy realization of recent latent unobserved new infections (incidence); denoted by I_u during time interval u.

3.2.2 Modeling Incidence

We first differentiate between incidence during the observation period, when case data are available, and incidence prior to the observation period. It is rare in practice to begin analysis of an infectious disease epidemic at the exact start of the epidemic. We follow Scott et. al. in modeling a number of unobserved incidence values (often called seeded incidence) drawn from a hierarchical exponential model [Scott et al., 2021, Bhatt et al., 2023]. That is, for

$t = -n, -n-1, \dots, 0,$

$$\lambda \sim \text{Exponential}(\eta),$$

$$I_t \sim \text{Exponential}(\lambda).$$

We model latent incidence during the observation period as a latent gamma random variable:

$$I_t \mid \mathbf{I}_{-n:t}, R_t \sim \text{gamma}\left(R_t \sum_{u=-n}^{t-1} g_{t-u} I_u \nu, \nu\right), t = 1, \dots, T,$$

where $\mathbf{I}_{-n:t}$ is the set of all previous incidences between times $-n$ and t, g_t is the discretized probability density function of the generation time (the time from an individual becoming infected to infecting someone else) distribution for the interval t, and R_t is the effective reproduction number at time interval t. Parameter ν, describing the proportional mean-variance relationship of the above gamma distribution, receives its own prior:

$$\log(\nu) \sim N(\mu_\nu, \sigma_\nu^2).$$

We assume g_t to be known. Svensson and Champredon et al. [Svensson, 2007, Champredon and Dushoff, 2015, Champredon et al., 2018] have highlighted that in a closed population, the generation time distribution depends on population dynamics, i.e., it changes over time depending on the number of susceptibles available, somewhat similarly to the effective reproduction number. This is not taken into account in our model (nor, to our knowledge, in any model in this class of estimators). Instead we use the intrinsic generation time distribution that assumes a fully susceptible population.

Note that under this model,

$$\mathrm{E}(I_t \mid \mathbf{I}_{-n:t}, , R_t) = R_t \sum_{u=-n}^{t-1} I_u g_{t-u}, \tag{1}$$

$$\mathrm{Var}(I_t \mid \mathbf{I}_{-n:t}, , R_t) = R_t \sum_{u=-n}^{t-1} I_u g_{t-u}/\nu. \tag{2}$$

The assumed mean relationship lies at the heart of branching process inspired methods for estimating the effective reproduction number [Fraser, 2007]. Pakkanen et al. [2023] show that Equation (1) is justified under a formulation of disease transmission modeled as a variation on the Crump-Mode-Jagers branching process. Regardless of the underlying model, we think it is beneficial to allow for incidence to change stochastically. To this end, we model incidence as an auto-regressive gamma process while preserving the branching process inspired mean model (1). By modeling incidence as a continuous random variable, we are able to use Hamiltonian Monte Carlo to approximate the posterior distribution of our model parameters. The mean-variance relationship of the gamma distribution is also somewhat convenient, as it allows for over-dispersion in the variance of incidence through parameter ν.

To allow for the effective reproduction number to change over time, we model it as a random walk on the log scale:

$$\log R_1 \sim \mathrm{Normal}(\mu_{r1}, \sigma_{r1}^2),$$

$$\log R_t | \log R_{t-1}, \mathbf{I}_{-n:t-1} \sim \mathrm{Normal}\left(\log R_{t-1}, \frac{\sigma^2}{T-1}\right), \quad t = 2, \dots, T.$$

The prior distribution of σ, $\log(\sigma) \sim N(\mu_\sigma, \sigma_\sigma^2)$, is chosen to reflect beliefs about the total amount of possible variation in the effective reproduction number over the course of the observed period.

3.2.3 Modeling Observed Cases

Depending on the context of an infectious disease, the relationship between observed cases and incidence can be complex. One challenge relates to testing supply. The number of cases observed is always a function of the number of diagnostic tests administered. In the context of a novel infectious disease, testing supplies may change rapidly as new technologies are developed, approved, and deployed. Thus, we model observed cases (O_t) conditioned on previous and current incidence ($I_t, \mathbf{I}_{-n:t-1},$) as a negative binomial random variable, where the mean of the negative binomial random variable is a function of incidence (as in [Scott et al., 2021, Bhatt et al., 2023, Abbott et al., 2020]), the number of tests administered, and a detection parameter ρ, with over-dispersion parameter κ:

$$\begin{aligned}
\kappa &\sim \text{Truncated-Normal}(\mu_\kappa, \sigma_\kappa^2),\\
\log \rho &\sim \text{Normal}(\mu_\rho, \sigma_\rho),\\
D_t &= \sum_{j=-n}^{t} I_j d_{t-j},\\
O_t \mid I_t, \mathbf{I}_{-n:t}, , \rho, \kappa, M_t &\sim \text{Neg-Binom}(\rho \times M_t \times D_t, \kappa), t = 1, \dots T,
\end{aligned} \tag{2}$$

where $\rho \times M_t \times D_t$ is the mean of the negative-binomial distribution. As defined above, M_t is the total number of diagnostic tests administered during time interval t. As a result, the detection rate for time t is $\rho \times M_t$, which allows the detection rate to change over time as a function of the number of tests available. With a detection rate which depends on tests, the model can discern between situations where cases increase because of increases in latent incidence, as opposed to increases in the number of tests administered. The weights d_{t-j} are discretized weights of the delay period distribution, that is, the time from infection to detection. Delays occur for a variety of reasons, based on when the difference between when individuals are infected and when they test, as well as delays in reporting the results of the test. In our study, we will use simulations and data where the only delay is caused by our assumption that cases represent individuals transitioning from the latent stage of infection to the infectious stage. Thus, for this chapter, d_{t-j} are discretized weights of the latent period distribution. Note that we allow for cases observed at time t to come from incidence observed at time t as well; this can be adjusted depending on how quickly a particular disease spreads and at what granularity observations are recorded.

It is difficult to choose generic priors for κ and ρ, as they both depend in some way on properties of the surveillance system used to collect data. We address this challenge in the sections below.

3.2.4 Prior for Case Over-Dispersion

In our experience, some choices of the prior distribution for κ result in poor Markov chain Monte Carlo (MCMC) convergence. To overcome this issue, we developed an approach for choosing the prior distribution for κ inspired by Empirical Bayes methods. We fit a Bayesian thin plate regression spline to the time series of cases, assuming a negative-binomial distribution with the mean number of cases being a nonparametricaly estimated function of time, then use the posterior estimate for the over-dispersion parameter to construct the prior for our model [Wood, 2017]. We use `brms` (version 2.15.0) to fit the regression spline to observed cases [Bürkner, 2017]. This method has drawbacks from a theoretical perspective, because the spline-based model is fit to the same data that is then analyzed with our semi-mechanistic model. For simulations, this is easily overcome by fitting the spline to a simulated data set that is then not analyzed by our model, this is the approach we took for our simulation study. For real data analysis, one solution is to fit a spline to data from an outbreak occurring in a similar location to the one being analyzed. For this chapter, we put aside theoretical concerns and fit a spline-based model to each real data set used in this chapter to derive the prior for κ and then applied our model. We choose the parameters of the prior by minimizing a squared loss function, searching for prior parameters that minimized the squared difference between the quantiles of the spline posterior, and the empirical quantiles of the candidate prior distribution.

3.2.5 Prior for the Case Detection Rate

Choosing the prior for the case detection parameter ρ likewise requires some care, because the meaning of ρ depends on the number of diagnostic tests in the data. We propose the following procedure: first construct a plausible range for what proportion of incidence has been observed. Then, using the the 50% quantile of tests in the observed test time series, construct a prior for ρ that matches the prior for the overall mean case detection rate. In practice, we can construct the prior for ρ using other quantiles as part of sensitivity analyses. For simulations we use a ρ prior derived using the 50% quantile, for real data anlysis, we use the 25% quantile which we found improved MCMC convergence.

3.2.6 Bayesian Inference

Let $\mathbf{R} = (R_1, R_2, \dots, R_T)$ denote the vector of effective reproduction numbers and $\mathbf{I} = (I_{-n}, \dots, I_T)$ the vector of latent incidence counts. We are interested in the posterior distribution of our model parameters:

$$P(\mathbf{I}, \mathbf{R}, \rho, \kappa, \nu, \lambda, \sigma \mid \mathbf{O}) \propto P(\mathbf{O} \mid \mathbf{I}, \rho, \kappa) P(\mathbf{I} \mid \mathbf{R}, \nu, \lambda) P(\mathbf{R} \mid \sigma) \pi(\rho, \kappa, \nu, \lambda, \sigma).$$

Here $P(\mathbf{O} \mid \mathbf{I}, \rho, \kappa)$ defines the emissions model, $P(\mathbf{I} \mid \mathbf{R}, \nu, \lambda)$ defines the latent case model, $P(\mathbf{R} \mid \sigma)$ the random walk prior for the effective reproduction number and $\pi(\rho, \kappa, \nu, \lambda, \sigma)$ the prior on all other model parameters.

We use Hamiltonian Monte Carlo, implemented in the R package `rstan` (version 2.21.2) to approximate the above posterior distribution [Stan Development Team, 2020]. For the remainder of this chapter we will refer to our effective reproduction number estimation method as Rt-estim-gamma.

3.2.7 State-of-the-Art Methods

`EpiEstim` models observed cases as incidence, and assumes that

$$I_t \mid I_1, \ldots, I_{t-1}, R_t \sim \text{Poisson}\left(R_t \sum_{u=1}^{t-1} I_u g_{t-u}\right).$$

To facilitate smooth estimates, the effective reproduction number is assumed to be fixed for a given period of time, and then repeatedly estimated for all such periods in the data set. We choose a period size of one week, and allow for an uncertain generation time, refitting the model using different values for g_{t-u} (see Cori et al. [2013] for details). The prior on the effective reproduction number for each window is a gamma distribution with shape parameter 1 and scale parameter 5.

Using R package `epidemia` (version 1.0.0) we created the Rt-estim-normal model. In this model, latent incidence is an autoregressive normal random variable with variance equal to the mean multiplied by an over-dispersion parameter so the mean-variance relationship is the same as in our autoregressive gamma model. We model cases as a negative-binomial random variable, using the latent period distribution as the delay distribution (though in `epidemia` it is assumed cases cannot be generated from the current latent incidence). The case detection prior is chosen to reflect a range of plausible values for case detection depending on the simulation scenario and real data. For observed cases, we attempted to use a prior for the over-dispersion parameter that had similar values to the prior used in our model for the over-dispersion parameter of the negative binomial distribution, but found this led to issues with MCMC convergence. As such, we use the default prior for the inverse of the over-dispersion parameter implemented in `epidemia`. All other priors used are default priors from the `epidemia` package. For a full description of Rt-estim-normal, see the Appendix.

3.2.8 `EpiEstim` as an Autoregressive Generalized Linear Model

Under the basic `EpiEstim` modeling framework, the only value in Equation 1 that is random is R_t. Consequently, `EpiEstim` can be mimicked via Poisson regression with an identity link and no intercept. This raises the possibility of assessing the presence of over-dispersion in case data using standard statistical methods. To be more explicit, we can rewrite Equation 1 in the style of a generalized linear model (GLM):

$$\mathrm{E}[I_t \mid \mathbf{I}_{-n:t}, , R_t] = \eta = \beta_1 x_1.$$

In this construction, $\beta_1 = R_t$ and $x_1 = \sum_{u=1}^{t-1} I_u g_{t-u}$ is the weighted sum of previous incidence. After choosing an arbitrary number of previous incidences to include in x_1, we can construct x_1 manually for every observed incidence at time t with the requisite number of observed previous incidences. To estimate the effective reproduction number over time, we use Poisson regression repeatedly on subsets of data, where each subset has observations equal to the length of the smoothing period used in `EpiEstim`. For example, we can implement Poisson regression on data sets with 4 observations, estimating a β_1 which is fixed for those 4 observations. This is equivalent to using a period of 4 in `EpiEstim`. We assign the estimated effective reproduction number to the last date among the 4 observations, and change the settings of `EpiEstim` to match its estimates to the last observation as well. In addition to mimicking `EpiEstim` with a Poisson GLM, we mimic `EpiEstim` using a quasi-Poisson GLM. Using the quasi-Poisson's estimated over-dispersion parameter, we can assess how well the assumed mean variance relationship of the Poisson GLM matches the empirical variance seen in the observed data. We implement both GLM versions of `EpiEstim` and compare to the simplest version of EpiEstim using a fixed generation time in order to motivate the use of more complex models. All code and data needed to reproduce the results are available on GitHub at https://github.com/igoldsteinh/improving_rt.

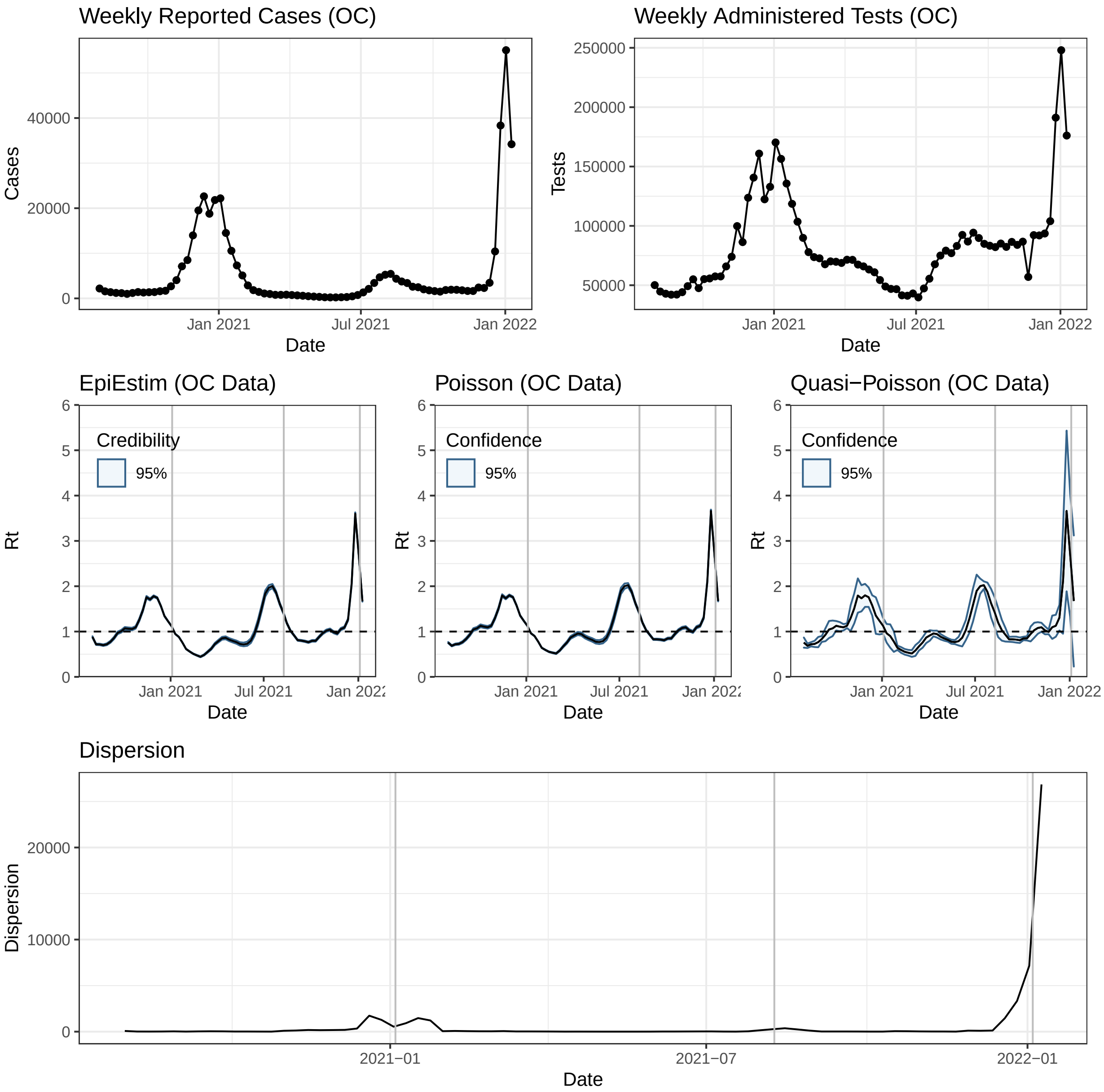


Figure 3.1: Estimation of the effective reproduction number of SARS-CoV-2 in Orange County, CA from Aug 2nd 2020 through January 15th 2022. The top row displays the observed cases and total diagnostic tests administered for the period. The middle row displays estimates of the effective reproduction number from `EpiEstim`, and from two GLM based mimics of `EpiEstim` using Poisson, and Quasi-Poisson regression. Blue regions represent 95% credible or Wald confidence intervals, while black lines represent posterior median or point estimates of the effective reproduction number. The final row displays the estimated over-dispersion parameter from the Quasi-Poisson regression model. Grey vertical lines mark the date maximum statewide cases were reported for the original winter 2020 wave, the summer 2021 wave, and the winter 2021 wave.

3.3 Results

3.3.1 GLM `EpiEstim` Applied to the SARS-CoV-2 Outbreak in Orange County, CA

To motivate the use of more complex models for estimating the effective reproduction number, we applied `EpiEstim` and our two GLM mimics of `EpiEstim` to case data from the SARS-CoV-2 outbreak in Orange County, CA from May 17th 2020 to January 15th 2021. We used a window size of 4 for `EpiEstim` and corresponding data sets with 4 observations for the GLM mimics. Data and effective reproduction number estimates are displayed in Figure 3.1. The Poisson GLM closely tracks the effective reproduction number trajectory estimated by `EpiEstim`. However, the Quasi-Poisson estimate of the effective reproduction number has much wider confidence intervals than the Poisson GLM. This is because the estimated over-dispersion parameter in the Quasi-Poisson model ranges from 1.01 to 26851.84. This shows that the Poisson model for incidences may be inadequate, resulting in overconfidence of R_t inference.

3.3.2 Simulation Protocol

Simulated data for this chapter were generated from a stochastic SEIR model in R (version 4.0.4) using the `stemr` package (version 0.2.0) [R Core Team, 2020, Fintzi et al., 2022]. SEIR models generate an infectious disease outbreak at a population level, with the population divided into four compartments: susceptible, exposed (infected but not yet infectious), infectious, and removed (neither infectious nor susceptible). The changes in these compartments are governed by rate parameters that depend on the populations in the compartments. In our simulations, the mean latent period was 4 days, the mean infectious period was 7.5 days. Daily case data were generated from transitions from the E to the I compartment on day t, using a fixed number of tests and a negative binomial distribution. For all simulations, ρ was set to be 9×10^{-5} and κ was set to be 5.

The basic reproduction number R_0 was given a fixed trajectory, leading to similar R_t trajectories for each realization of the simulation. More details on the stochastic SEIR model used for simulation are available in Appendix section A.1.1. Note that the SEIR models used for the simulations do not match any of the models used for inference of the R_t trajectories. In other words, all our simulation results are produced in the presence of model misspecification — a desirable feature for a realistic simulation protocol.

We simulated three separate scenarios lasting 28 weeks, where all parameters were the same except for the number of tests at each time step. In Scenario 1, weekly tests were drawn from a normal distribution with parameters that remained constant over time. In Scenario 2, tests were held constant for the first six weeks of the simulation, then increased at varying rates over the next eleven weeks of the simulation. Scenario 3 was similar to Scenario 2, except that testing was held constant for the first eight weeks, and increased more quickly than in Scenario 2. All simulations were done on a daily time scale, then aggregated into weeks for analysis. The true effective reproduction number for a single week was taken to be the true effective reproduction number of the third day of that week. In all simulations, the first 11 weeks were not analyzed, leaving 17 weeks of data for analysis. For each scenario, we generated 100 simulations. Realizations of all three simulations are displayed in Figure 3.2.

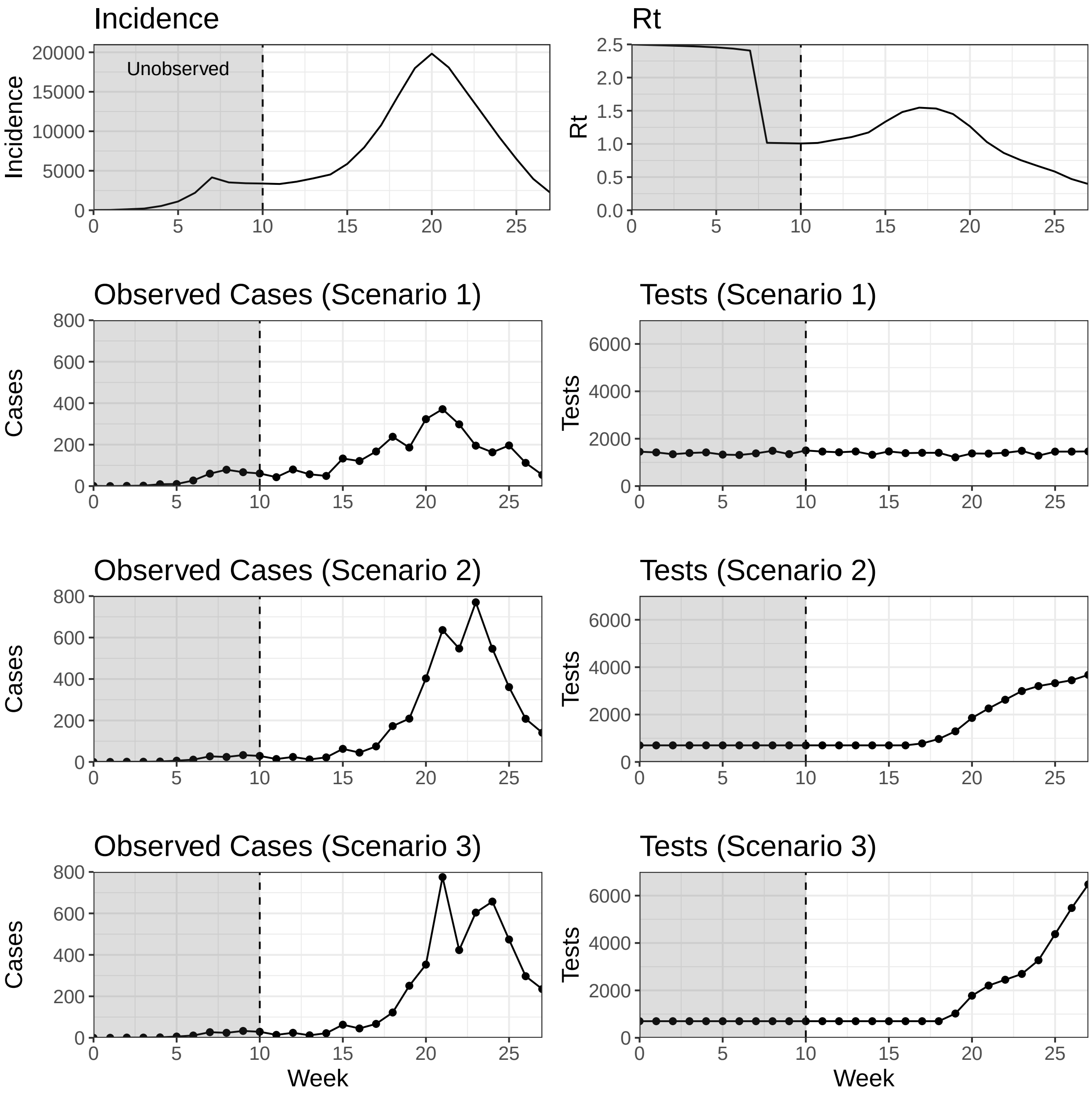


Figure 3.2: Simulated epidemic data generated from an SEIR model. Cases are generated using an emissions model that includes total diagnostic tests administered as a covariate. Three different testing scenarios are considered, all with the same underlying R0 trajectory. Included are underlying incidence and effective reproduction number trajectories. While these may vary slightly across simulations, they will be very similar due to identical infectious disease dynamics. In Scenario 1, tests are randomly sampled from a normal distribution. In Scenarios 2 and 3, tests stay flat and then increase at varying rates. Simulated epidemics start with 10 individuals and last for 28 weeks. The first 11 weeks are discarded, and are not used when simulated data are analyzed by the three effective reproduction number estimation methods.

Table 3.1: Priors used by the Rt-estim-gamma method in the simulation study.

Parameter	Simulation	Prior	Prior Median (95% Interval)
ν	All	Log-normal(-2, 0.7)	0.15 (0.03, 0.53)
σ	All	Log-normal(-0.66, 0.6)	0.52 (0.16, 1.68)
λ	All	Exponential(0.3)	2.31 (0.08, 12.26)
$\log R_1$	All	Normal(0, 0.75)	0.01 (-1.49, 1.49)
ρ	Scenario 1	Log-normal(-11.06, 0.3)	1.57E-5 (8.756E-6, 2.85E-5)
ρ	Scenario 2	Log-normal(-11.43, 0.3)	1.09E-5 (5.96E-6, 1.96E-5)
ρ	Scenario 3	Log-normal(-11.56, 0.3)	1.57E-5 (8.81E-6, 2.83E-5)
κ	Scenario 1	Truncated-Normal(59, 60)	72.00 (5.00, 183.15)
κ	Scenario 2	Truncated-Normal(33, 25)	35.65 (3.14, 83.23)
κ	Scenario 3	Truncated-Normal(70, 80)	88.84 (6.00, 235.41)

3.3.3 Simulation Results

For each model fit using `rstan`, we sampled 2000 posterior draws, discarding the first half as burn-in. Figure 3.3 visualizes the estimates for the effective reproduction number from `EpiEstim`, Rt-estim-normal and Rt-estim-gamma for the three data sets visualized in Figure 3.2. We checked convergence diagnostics for Rt-estim-normal and Rt-estim-gamma for all simulations and ensured adequate convergence of all models. More details are in the Appendix section A.1.4. Since `EpiEstim` does not provide estimates for the first time point in the series, we report only time points for which all three methods have estimates.

Credible intervals for `EpiEstim` frequently miss the true R_t values (covering between 6 and 9 of the 16 true values), while credible intervals for Rt-estim-normal and Rt-estim-gamma cover most true values across simulations. However, Rt-estim-gamma covers more true values than Rt-estim-normal, with narrower credible intervals (ranging between 11 and 16 values for Rt-estim-normal, and 16 values for every scenario for Rt-estim-gamma).

Figure 3.4 visualizes estimates of latent incidence from Rt-estim-normal and Rt-estim-gamma for the three data sets visualized in Figure 3.2. Rt-estim-normal credible intervals rarely cover the true incidence (covering from 0 to 5 to true values), while Rt-estim-gamma credible intervals generally do (covering 11 to 16 true values).

Posterior predictive distributions for cases for both Rt-estim-normal and Rt-estim-gamma are displayed in Appendix Figure A.2 (the posterior predictive distribution for `EpiEstim` is not readily available). For all three scenarios, for both models, 95% credible intervals from the posterior predictive distributions cover all observed data points. Rt-estim-gamma had generally narrower credible intervals than Rt-estim-normal.

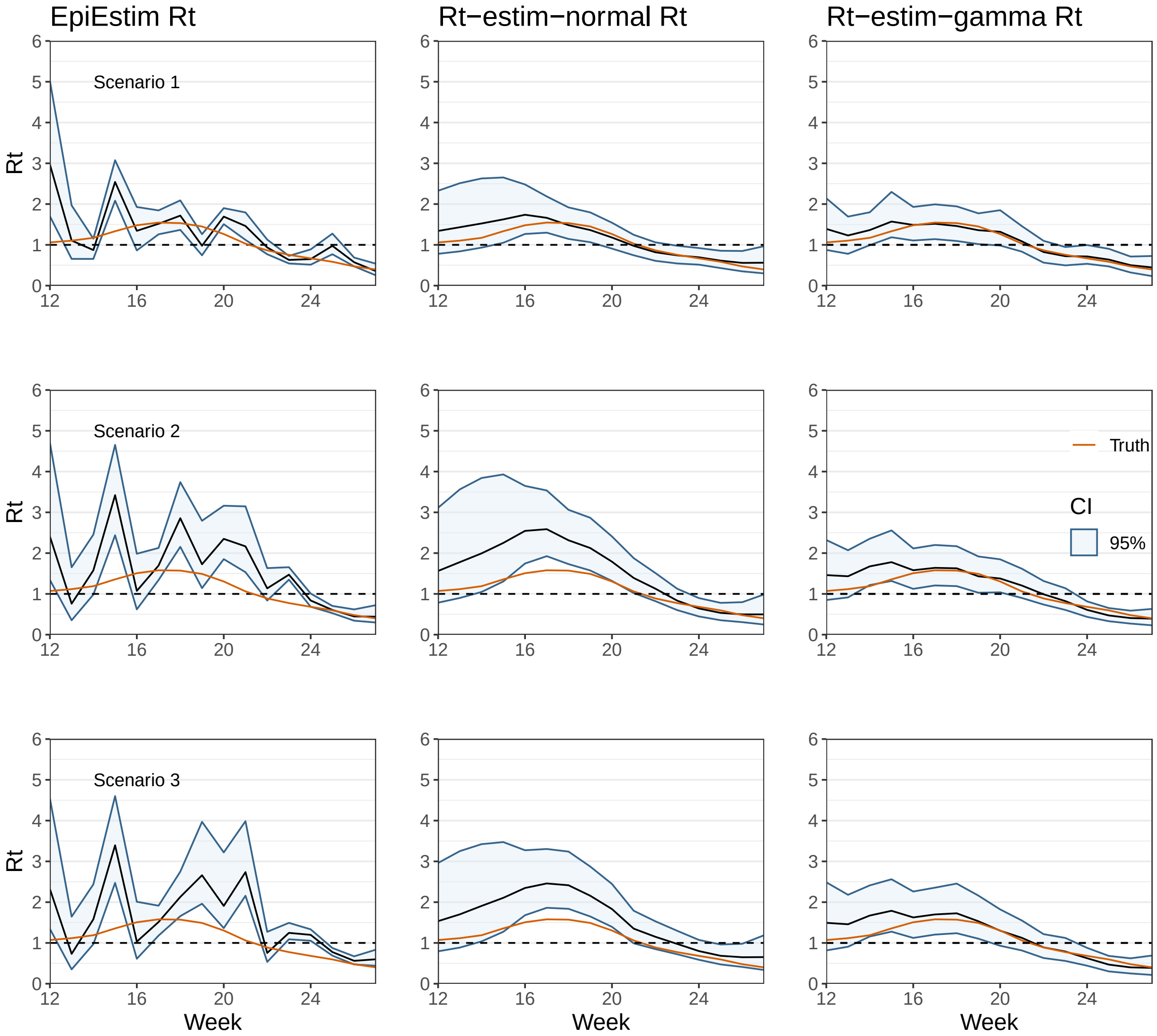


Figure 3.3: R_t estimation using three different methods for three simulated data sets under different testing scenarios. True R_t trajectories are colored in red, black lines represent median estimates from the posterior distribution, blue shaded areas are 95% credible intervals.

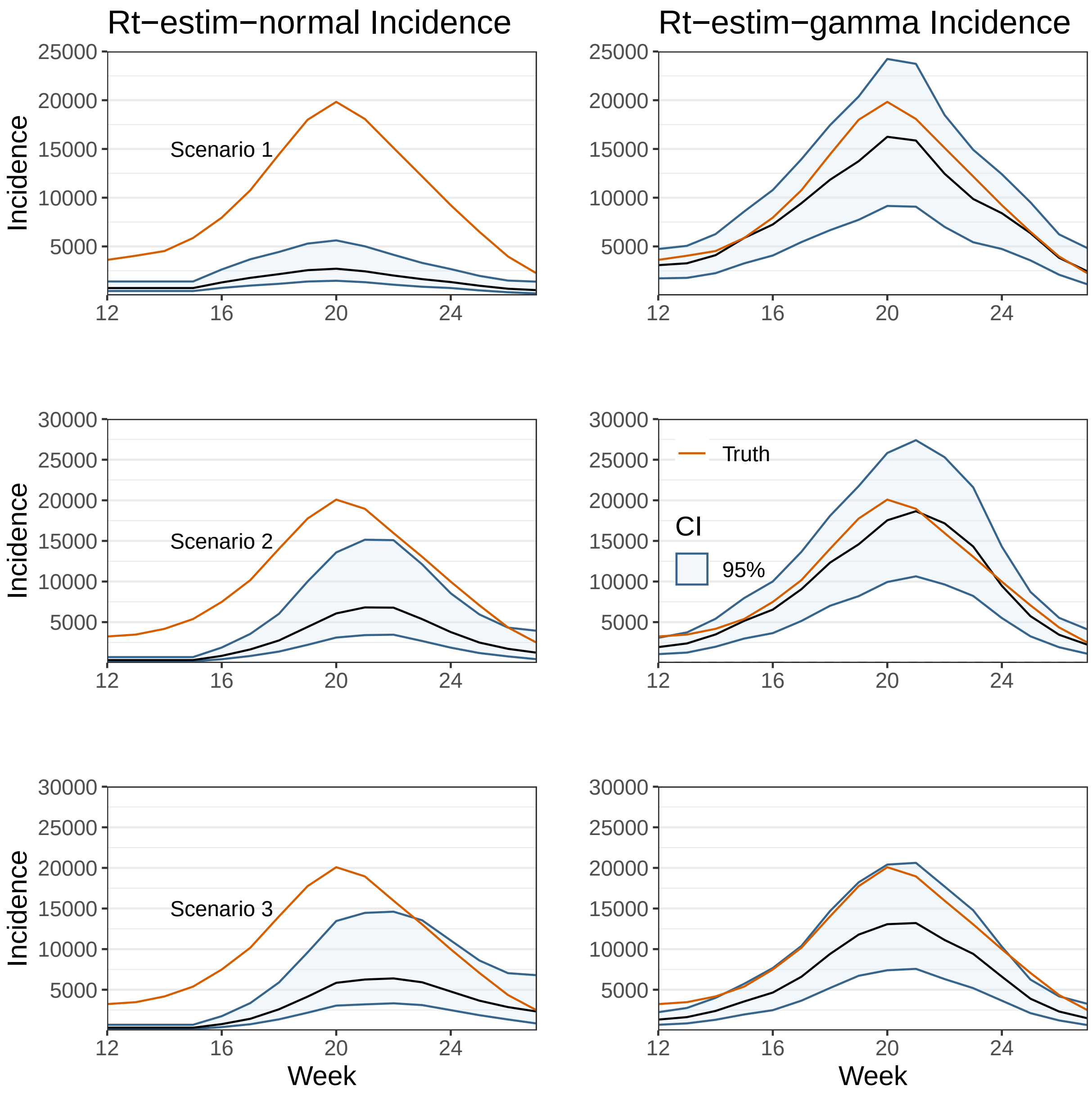


Figure 3.4: Incidence estimation using two different R_t estimation methods for three simulated data sets under different testing scenarios. True incidence trajectories are colored in red, black lines represent median estimates from the posterior distribution, blue shaded ares are 95% credible intervals.

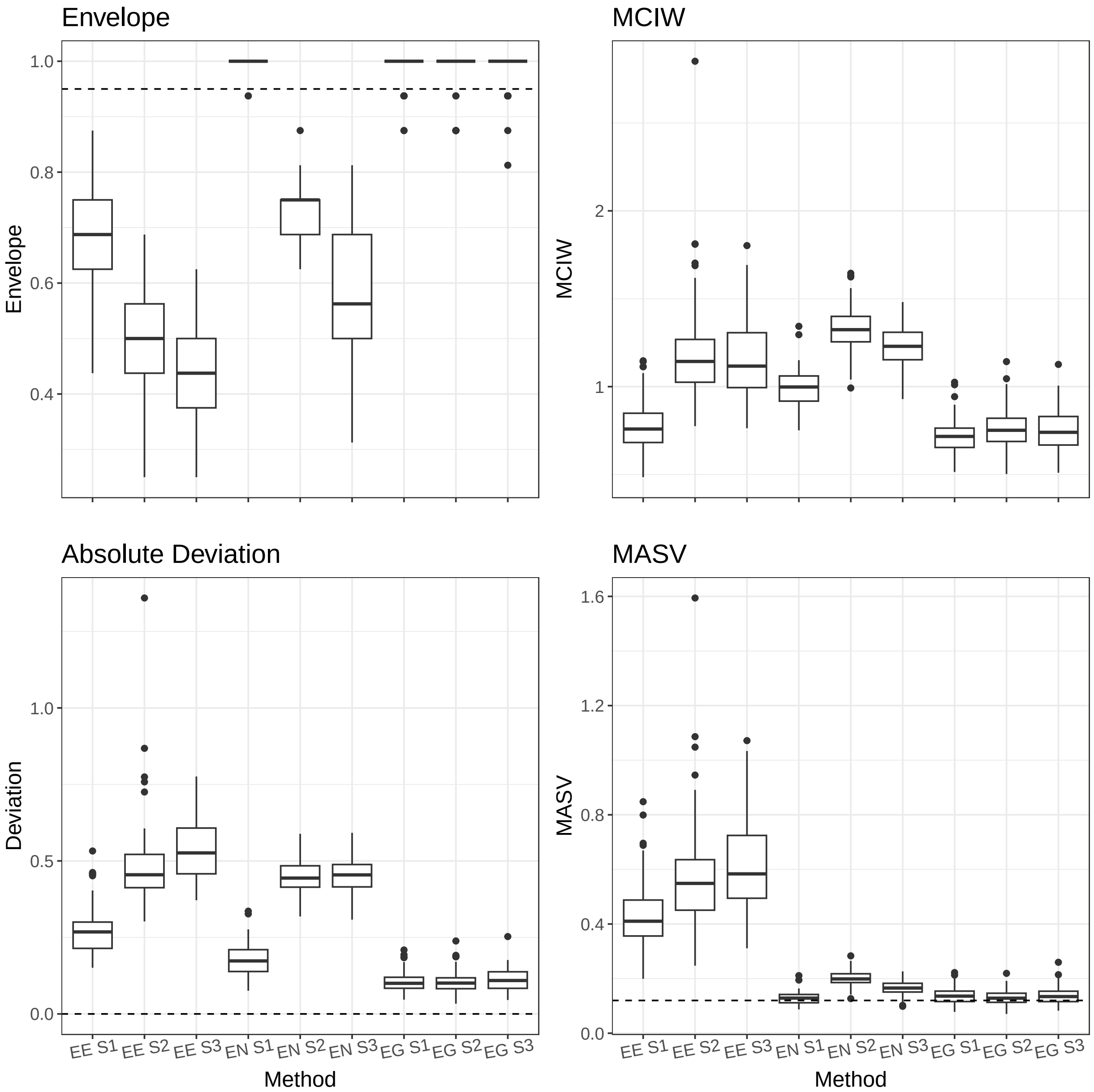


Figure 3.5: Frequentist metrics for `EpiEstim`, Rt-estim-normal, and Rt-estim-gamma applied to three different simulated epidemics. Envelope is a measure of coverage, taking the average coverage of 95% intervals over the time series. MCIW is the average mean credible interval width. Absolute deviation is the mean of the absolute value of the difference between the median R_t at each time point and the true value. Mean absolute standard deviation is the difference between the current median point estimate for R_t and the previous point estimate for R_t. The dashed line in the bottom right panel represents the true absolute standard deviation. Solid lines represent medians, hinges are upper and lower quartiles and whiskers are at most 1.5 times the inter-quartile range from the median. `EpiEstim` is denoted EE, Rt-estim-normal EN, and Rt-estim-gamma EG, while scenarios are marked S1, S2 and S3. EE S1 describes results of using `EpiEstim` in simulation scenario 1.

Because we are using a stochastic SEIR model to generate simulations, each simulation has a different, though similar in shape, true effective reproduction number curve (despite having the same true basic reproduction number curve). The range of true effective reproduction number curves is visualized in Figure A.1. We report frequentist metrics in order to summarise performance across a variety of different epidemic curves. Model performance on simulated data sets for each of the three models is summarized in Figure 3.5. For each metric, we summarize results in boxplots where solid lines represent medians, hinges are upper and lower quartiles and whiskers are at most 1.5 times the inter-quartile range from the median. Envelope is a measure of coverage. For each simulation the envelope is the proportion of time points for which a 95% credible interval from the posterior distribution captured the true value of interest. Mean credible interval width (MCIW) is the mean of credible interval widths across time points within a simulation. Absolute deviation is a measure of bias, and is the mean of the absolute difference between the posterior median and the true value at each time point. Finally, mean absolute sequential variation (MASV) measures how well each method captured the variation in the effective reproduction number across time by computing the mean of the absolute difference between the posterior median at time point t and the posterior median at time point $t-1$. We compare this to the true mean absolute sequential variation in each simulation. `EpiEstim` had the lowest envelope in all simulation scenarios. Rt-estim-normal had high envelope in Scenario 1 but dropped to lower values in scenarios with time-varying testing supply. Rt-estim-normal had the largest MCIW in all three scenarios, while Rt-estim-gamma had the smallest MCIW in all three scenarios. `EpiEstim` and Rt-estim-normal had relatively similar values for absolute devaition, Rt-estim-gamma had the smallest absolute deviation in all scenarios. Finally, `EpiEstim` had the largest MASV in all scenarios, while Rt-estim-normal and Rt-estim-gamma had relatively comparable MASV. For two of three scenarios, Rt-estim-gamma was closer to the true MASV than Rt-estim-normal. We ran three additional experiments using the data sets from Scenario 3 to better understand our model. All results are displayed in Appendix Figure A.4, with the results from Figure 3.5 included as a baseline comparison. In the first experiment, we halved each parameter in the hypo-exponential distribution and refit the model to the data sets from Scenario 3. This led to narrower credible intervals and lower envelope (see Appendix Figure A.4 for results). In the second experiment, we used a spline fit to the same data being analyzed in order to choose a prior for κ. We found no meaningful difference in performance. In the third experiment, we used a prior for ρ derived from the 25% quantile of tests, rather than the 50% quantile, this again led to no meaningful difference in performance with regards to estimating the effective reproduction number, though we expect it to change estimates of incidence. Overall, we find that our Rt-estim-gamma model outperforms the `EpiEstim` and Rt-estim-normal in all metrics, surprisingly even in Scenario 1, where the number of diagnostics tests did not vary appreciably over time (Rt-estim-gamma and Rt-estim-normal have similarly high envelope values in this case).

3.3.4 Estimating the Effective Reproduction Number of SARS-CoV-2 in Fifteen California Counties

We analyzed SARS-CoV-2 reported case data from fifteen California counties representing Northern California (Alameda, Sacramento, San Francisco, Santa Clara, Contra Costa), Central California (Fresno, Merced, Monterey, Stanislaus, Tulare) and Southern California (Los Angeles, Orange, Riverside, San Bernardino, San Diego). These counties represent more than 75 percent of the population of California, and differ widely along demographic, economic, and political characteristics. We analyzed data from August 2nd 2020 through January 15th 2022. Data are publicly available from the California Open Data Portal [California Open Data Portal, 2023]. Positive cases are associated with the date of their test, rather than the date they were reported.

To estimate the effective reproduction number of SARS-CoV-2 we must choose a generation time distribution to use in our models. Estimating intrinsic generation times from observed data is non-trivial Park et al. [2021]. Early efforts from Ferretti et al. [2020] and Ganyani et al. [2020] estimate the mean generation time to be between 5.5 and 5.2 days respectively. A more recent estimate of the mean intrinsic generation interval for the original version of SARS-CoV-2 estimated it to be 9.7 days [Sender et al., 2021], but the issue of optimal generation time inference seems far from settled. An additional complication is that a number of important variants of SARS-CoV-2 have spread over the course of the pandemic, and the generation times for the variants may differ from that of the original viral strain. Hart et al. [2022] found it is likely that the intrinsic mean generation time of the delta variant is shorter than that of the alpha variant, likewise a preliminary study by Abbott et al. suggests the intrinsic mean generation time of the omicron variant is shorter than that of the delta variant [Abbott et al., 2022]. We find the methodology of Sender et al. somewhat persuasive, and use their point estimate of the generation time (a log-normal distribution with mean 9.7 days) as the default generation time for the original SARS-CoV-2 strain and alpha variant versions of SARS-CoV-2. We compare these default findings to results using the Ferretti et al. point estimate distribution (a Weibull distribution with mean 5.5 days) in Appendix A.6.

We then created an alternative version of our model which allowed for changing the generation time distribution due to the delta and omicron variants. We changed the generation time distribution starting in July 2021, reflecting our assumption that delta variant dominated new cases by this point, and changed it again in December 2021, reflecting the same assumption about the omicron variant. Hart estimates the median reduction in the mean generation time for the delta variant is 15% as compared to alpha (we assumed alpha and wild-type had the same generation time) [Hart et al., 2022], while Abbott estimates the median reduction in the mean generation time for omicron is 28% as compared to delta [Abbott et al., 2022]. We created generation time distributions for these variants by minimizing a squared loss function to search for parameters such that the new distributions had the appropriate new mean generation time, while preserving the standard deviation of the original distribution (see Appendix for complete details). We tested whether this new model was needed by calculating the Bayes factor of the two models using data from Alameda County the `bridgesampling` package in R [Meng and Wong, 1996, Gronau et al., 2020], running both models for 26,000 iterations with the first 1000 iterations discarded as burn in on 3 chains. The point estimates for the marginal likelihood had error of 7% for the constant generation time model and 6% for the varying generation time model, with a reported Bayes Factor of 1.58 in favor of the model with variant-specific generation times. Even accounting for the margin of error, it is hard to conclude the varying generation time model was decisively superior to the constant generation time model, so we used the constant generation time model in this paper. Because we were testing a characteristic of the infectious disease that should generalize across locations, and because of the computational cost involved, we did not calculate Bayes factors for all fifteen counties.

Finally, we used the point estimate of the latent period distribution from Xin et al. [2022] as the delay distribution in our model, with a mean latent period of 5.5 days using a gamma distribution. For the alternative analysis using the Ferretti et al. distribution, we scaled this distribution by 0.5 to halve the mean latent period.

We fit `EpiEstim`, Rt-estim-normal (using the priors from the simulations), and Rt-estim-gamma (see Appendix for priors) to this data. The posterior summaries for the effective reproduction number as calculated by `EpiEstim` are displayed in Figure A.5 and those calculated by Rt-estim-gamma are displayed in Figure 3.6. Accompanying incidence posterior distributions and case posterior predictive distributions for Rt-estim-gamma are displayed in Figures A.6 and A.7 respectively. Visualizations of the priors and posteriors for non time-varying parameters for Rt-estim-gamma fit to Los Angeles County data are displayed in Figure A.14. After running into convergence issues with Rt-estim-normal, we reduced the data set to August 2nd 2020 through November 6th 2021 and fit Rt-estim-normal to this data set. The Rt-estim-normal results were generated using R version 4.2.2.

Priors for Rt-estim-gamma were the same as in the simulations, except that the prior σ had a mean of -0.61 (the range of plausible values was similar), and the priors for ρ and κ were chosen for each county individually using the protocols described in the methods section. We assumed the overall median proportion of observed incidence was 0.066. An example of posterior predictive intervals from the thin plate spline used to choose the prior for κ and from Rt-estim-gamma fit to SARS-CoV-2 case data from Alameda County, California, are visualized in Figure A.3.

Comparisons with Rt-estim-normal are displayed in figures A.8, A.9 and A.10. The Rt-estim-gamma results were generated in R version 4.1.2, but all packages were the same as those used to generate simulation results except for `Rcpp` which was version 1.0.8 rather than version 1.0.7. Overall, Rt-estim-gamma estimates were smoother and more uncertain than estimates from `EpiEstim`, but less smooth and uncertain than those produced by Rt-estim-normal. This behavior is consistent with model performance in the simulation scenarios. Rt-estim-gamma estimates tended to estimate less extreme magnitudes than Rt-estim-normal estimates, and while the two models produced broadly similar estimates of the trajectory of the effective reproduction number, they differed in some counties in significant ways. For example, in San Diego county, the median estimate from Rt-estim-normal is always above 1 before January 2021, while the median estimate from Rt-estim-gamma is below 1 for parts of this period. Additionally, in all counties the median estimate from Rt-estim-gamma crossed below 1 before the median estimate for Rt-estim-normal in fall 2021. Rt-estim-normal and Rt-estim-gamma produced different estimates of the latent incidence (Figure A.9), but both produced 95% posterior predictive intervals for the observed cases which had good coverage in all counties (Figure A.10).

The results using Rt-estim-gamma with a mean generation time of 5.5 days are displayed in Figures A.11, A.12 and A.13. Using a shorter generation time led to generally smaller estimates of the peak effective reproduction number with narrower credible intervals. However, the trajectories using either generation time were similar, and the estimated trajectories agreed on when the median reproduction number was above or below one.

Median estimates for the effective reproduction number were larger during the summer 2021 wave than during the winter 2020 wave. The estimate of the reproduction number during the winter 2021 wave was similar to that of summer 2021 wave except in a few counties where it was larger, such as Los Angeles and Alameda counties. Trajectories were similar across counties, but varied in timing and magnitude from county to county. For instance, the peak reproduction number in the Winter 2020 wave was estimated to occur in the week of November 1st in Sacramento County, and the week of November 22nd in Los Angeles County.

3.4 Discussion

We presented a model for estimating the effective reproduction number using time series of observed cases and diagnostic tests, as well as methods for choosing key priors for the model. We tested the model on simulated data sets, showing it can successfully estimate the true effective reproduction number when data are generated from a stochastic compartmental model. We also tested other models used for estimating the effective reproduction number, demonstrating that when testing supply is relatively constant, a case observation model that ignores testing is reasonable, but when testing supply changes rapidly, ignoring testing leads to poor model performance. Using data from the SARS-CoV-2 epidemic in California, we have shown how using a model fit to case observation data that incorporates testing data leads to different conclusions about the trajectory and magnitude of the effective reproduction number in real world epidemics.

We found that `EpiEstim` had poor performance across all simulation scenarios. In contrast, an assessment of EpiEstim by Gostic et. al found it had reasonable performance on simulated data and recommended it over other existing methodologies (models available in epidemia were not assessed in this chapter) [Gostic et al., 2020]. However, Gostic et. al. only tested EpiEstim on simulated data sets where the true incidence was known. In our study, we tested EpiEstim on data sets where cases were noisy realizations of unobserved incidence, a much more realistic scenario for many diseases, such as SARS-CoV-2. The performance of our GLM versions of EpiEstim on data from the SARS-CoV-2 epidemic in Orange County, CA provides one reason for this poor performance. Modeling reported cases as a Poisson random variable assumes a stringent mean-variance relationship that is likely to under-estimate uncertainty. We do not recommend using EpiEstim to estimate the effective reproduction number when there is reason to believe reported cases do not reflect true incidence.

When testing was relatively constant, the Rt-estim-normal model, which assumes latent incidence but ignores tests, still performed well. Even in this scenario, the Rt-estim-gamma model we developed for this chapter had smaller mean credible interval widths and smaller absolute deviations. This suggests that our other modeling choices beyond including tests as a covariate, such as the use of the gamma distribution to model latent incidence, and our process for choosing the prior for the case over-dispersion parameter, had positive effects on model performance. As the SARS-CoV-2 pandemic has unfolded, a number of modeling groups have developed similar techniques for estimating the effective reproduction number. We have demonstrated how modifying distributional assumptions and developing protocols for choosing priors can have significant impact on model performance. We hope these findings motivate the larger community of researchers focused on modeling the effective reproduction number to revisit their work and establish best practices for this class of models.

In simulation scenarios where testing supply increased dramatically, we were still able to successfully estimate the effective reproduction number using the Rt-estim-gamma model. Our findings suggest incorporating testing data is a viable strategy for using case data, which should improve the accuracy of efforts at now-casting the effective reproduction number. It is worth noting that we avoided using a delay distribution in Rt-estim-gamma which incorporated reporting delays, instead using data where cases and tests were tied to the date of the test. This should not prevent the use of Rt-estim-gamma for up-to-date now-casting even though counts of the most recent cases and tests will inevitably be under-counts, so long as the proportion of positive to total tests is independent of reporting delays. We assumed this was the case when applying Rt-estim-gamma to the SARS-CoV-2 data from California. If this assumption proves not to be true, then approaches that do not use testing data and incorporate more elaborate delay distributions, such as those of Abbott et al. [2020] and Bhatt et al. [2023] are probably a better choice. Similarly, because our model relies on the proportion of positive to total tests, rather than the raw counts of positive tests, it should be robust to changes in types of tests available, so long as reported positive cases used the same kinds of tests recorded in total diagnostic tests. This allows us to avoid any problems that arise from the availability of rapid tests for SARS-CoV-2 during the omicron wave.

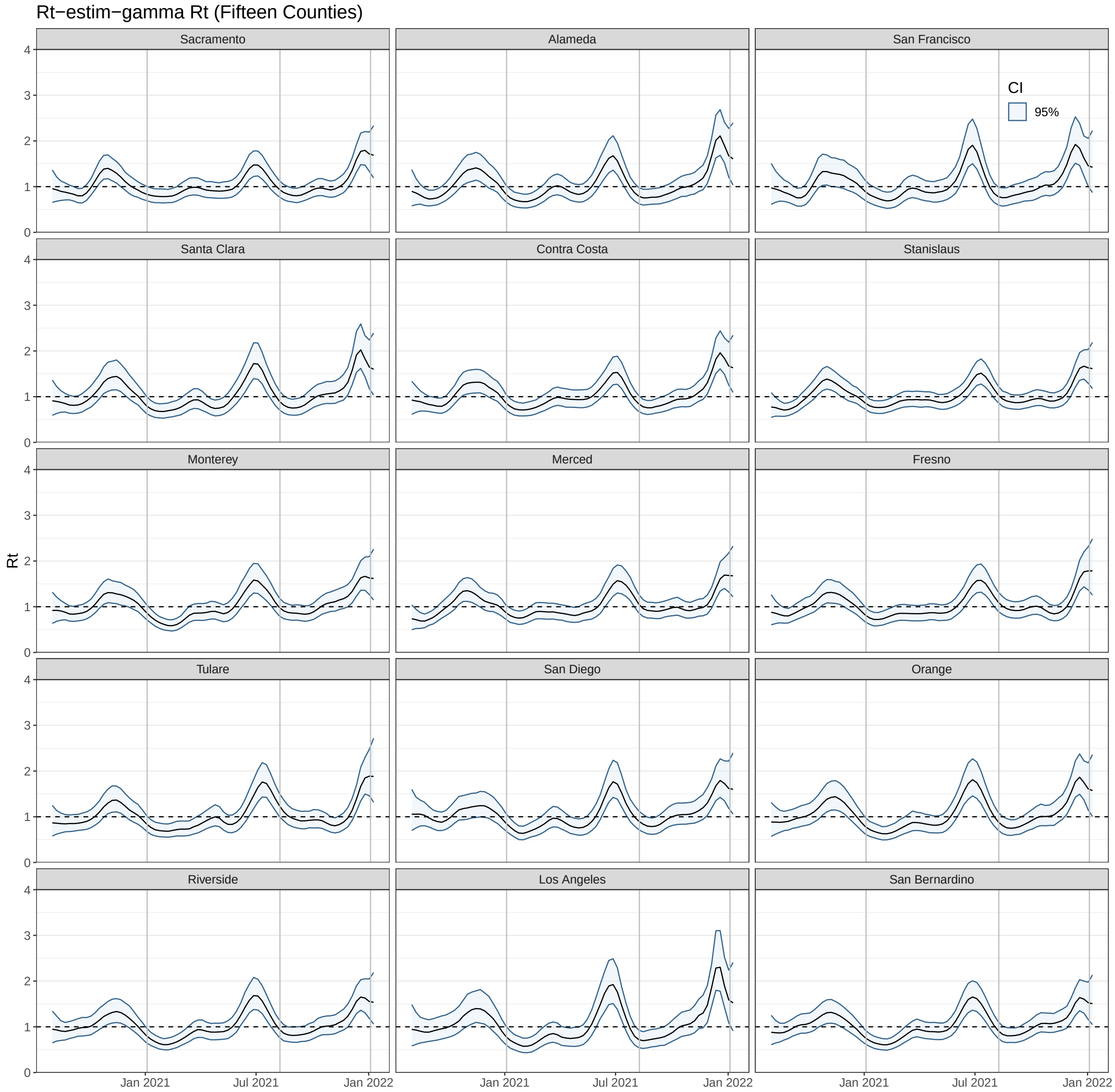


Figure 3.6: Estimates of the effective reproduction number of SARS-CoV-2 from Rt-estim-gamma applied to fifteen counties in California, USA from August 2nd 2020 through January 15th 2022. Blue shaded regions are 95% posterior credible intervals. Black lines are medians. Grey vertical lines mark the date maximum statewide cases were reported for the original winter 2020 wave, the summer 2021 wave, and the winter 2021 wave.

In a representative set of simulations, even when Rt-estim-normal posteriors captured the effective reproduction number, its posterior estimates for latent incidence did not capture the true latent incidence. To a lesser extent, the same was true of the Rt-estim-gamma model.

We have not yet seen any discussion as to the accuracy of incidence estimation for this class of models. Our findings suggest incidence estimates should not be trusted, as there are many values for incidence that lead to the same observed cases and the same reproduction number estimates. Estimates where a population size are taken into account, such as in Mishra et al. [2020], may be more trustworthy, but we recommend running a simulation study first to verify this.

One important limitation of our method is that we condition on the number of tests and use them as a covariate, rather than modeling them jointly with cases. We would expect in practice that the number of tests is also a function of past incidence (with cases rising, more tests will be administered). In mathematical terms, a joint model of cases and tests could be written as

$$P(\mathbf{O}, \mathbf{M} \mid \mathbf{I}, \rho, \kappa) = P(\mathbf{O} \mid \mathbf{M}, \mathbf{I}, \rho, \kappa) P(\mathbf{M} \mid \mathbf{I}, \rho, \kappa).$$

Our method only uses the first term of this product. This leaves our method open to potential bias from model misspecification. In the simpler context of regression without latent variables, this issue is sometimes called “feedback”, a thorough treatment of the topic is available in Chapter 12 of Diggle et al. [2002]. While we think that in practice this will not be a concern in situations where cases and tests increase and decrease together in response to changes in incidence, the possibility does exist. For instance, suppose the testing policy during the peak of an epidemic was that individuals with symptoms could not be tested, as anyone with symptoms should simply assume they have been infected. Tests might still increase in response to increased incidence from the wave, but cases could decline, because no symptomatic people were testing. In such a scenario, we would expect our model to fail. Modeling tests is a non-trivial problem, and implementing a joint model of cases and tests is a promising future direction.

In this paper, we used gamma densities to in order to model changes in latent incidence stochastically. While our choice of a gamma distribution has some desirable benefits, namely that it allows us to use HMC to generate posterior samples, and that it allows for overdispersion in the variance, there is definitely room for improvement in modeling latent incidence. Recent work by Penn et al. [2022] provides an interesting avenue for improvement, with explicit calculations of the variance of the transition distributions of a time-varying general branching process.

In the real data analysis, we used Rt-estim-gamma with a prior for the over-dispersion parameter derived from a spline fit to the same data as Rt-estim-gamma. This is a workaround we developed to avoid computational problems related to using Hamiltonian Monte Carlo when the prior for the over-dispersion parameter strongly conflicts with the data. Another MCMC method, such as Zig-Zag sampling [Bierkens and Roberts, 2017, Corbella et al., 2022b], may not have this issue, and so we could avoid this procedure. While not ideal, we tested our model using this procedure for choosing the over-dispersion prior on simulated data, and found no discernible loss in performance.

It is clear that more sophisticated representations of the generation time distribution which could change according population dynamics could be incorporated into our model. While this might lead to improved model performance, it is encouraging that in simulations, our model performed well despite using a fixed generation time. It is equally encouraging that our experiments on both simulated and real data showed our model was reasonably robust to different generation time distributions.

One obvious area for improvement in this space is allowing the prior on case detection (ρ) to change over time to better reflect changes in testing policy. For instance, at the start of the SARS-CoV-2 pandemic, only symptomatic individuals could get tested in California, whereas in Fall 2021, anyone was eligible to receive a test. Sherratt et al. [2021] also highlighted how changes in testing eligibility may result in estimating spurious changes in the effective reproduction number. We have found case data alone is insufficient to make a time varying detection parameter identifiable. Incorporating other sources of data that facilitate real-time estimation, such as data from wastewater treatment facilities, may enable models with time varying case detection parameters. Enabling effective reproduction number estimation methods to incorporate multiple data streams seems like a fruitful area of future research.

Chapter 4

Semiparametric Inference of Effective Reproduction Number Dynamics from Wastewater Pathogen Surveillance Data

4.1 Introduction

For many pathogens, infected individuals will shed copies of the pathogen through fecal matter over the course of their infection. Viral gene concentrations measured in wastewater samples are a noisy aggregate of the concentrations of genomes generated by infected individuals connected to the wastewater system, and thus provide insight into the dynamics of the spread of an infectious disease [Hillary et al., 2020, Polo et al., 2020]. One promising use for pathogen genome concentrations as data is estimation of the effective reproduction number. The effective reproduction number (R_t), the average number of individuals an infectious person at time t would subsequently infect, is a useful way of describing the state of an infectious disease epidemic. When R_t is below 1, we expect the number of new infections to decrease; the reverse is true when R_t is above 1. In this paper, we develop a new method for estimating the effective reproduction number from pathogen genome concentrations collected from wastewater, and as a bi-product, show how this method can be used to estimate the effective reproduction number from case data as well.

Recently, a number of studies have evaluated SARS-CoV-2 (the causative agent of COVID-19) RNA concentrations measured in wastewater as a potential data source, comparing them to both prevalence counts, counts of reported cases, and case rates [Morvan et al., 2022, Acer et al., 2022, Song et al., 2021, Zhan et al., 2022, Zulli et al., 2021]. Wade et al. [2022] provide a useful introduction to the many sources of uncertainty in the pathogen genome concentration data generation process.

From the perspective of inferential methods, pathogen genome concentrations are potentially less biased data than counts of new cases, which can be biased by policies regarding testing availability and the willingness of the population to test [Li et al., 2020]. To our knowledge, there have been relatively few attempts to use pathogen genome concentrations collected from wastewater (henceforth referred to as wastewater data) to estimate the effective reproduction number. Huisman et al. [2022b] adapted their case based method [Huisman et al., 2022a] and used SARS-CoV-2 wastewater data to create a synthetic time series of hypothetical case data that is then analyzed with the widely used case-based method `EpiEstim` [Cori et al., 2013]. While easy to use, the synthetic incidence is truncated on the assumption that wastewater data observed in the present contains little information about the number of newly infected individuals. As a consequence, the final estimate of the effective reproduction number is truncated as well. Nourbakhsh et al. [2022] used a classic compartmental model where RNA concentrations were modeled as noisy realizations of the number of currently infectious and recently recovered individuals. While the model produces inference on a number of parameters beyond the effective reproduction number, it also requires a number of parameters to be specified by users in order to produce inference, many of which are difficult to verify in practice.

Taking inspiration from previous work on non-parametric modeling of the transmission rate in compartmental models [Xu et al., 2016] and birth-death modeling in infectious disease phylodynamics [Stadler et al., 2013], we introduce a simpler compartmental model with only compartments needed to estimate the effective reproduction number and equip it with a Bayesian nonparametric inference framework. This simpler model combined with our Bayesian nonparametric framework lets us avoid some difficult-to-verify assumptions, while still estimating the effective reproduction number from wastewater data.

In this paper, we first introduce the classic compartmental modeling framework, then our new wastewater-based method for estimating the effective reproduction number. We test our new model against compartmental models fit to case and wastewater data as well as a state-of-the-art wastewater-based method on simulated data. Finally, we apply our new method to estimate the effective reproduction number of SARS-CoV-2 in Los Angeles, California using SARS-CoV-2 RNA concentrations collected from a large wastewater treatment plant.

4.2 Methods

4.2.1 Available Data

We will consider two types of surveillance data; concentrations of pathogen genomes measured in wastewater, and reported new cases, observed at times $t_1, \dots t_T$. It is common practice to measure the concentration from the same sample of wastewater multiple times, producing multiple measurements called replicates. In real world data sets, an average

Table 4.1: Parameters of the SEIR model.

Parameter	Interpretation
β_t	time-varying transmission rate
$1/\gamma$	average time infected but not infectious (average length of the latent period)
$1/\nu$	average length of the infectious period
N	total population size

of replicates is often reported. We will consider both raw concentrations and averages in this chapter. We define $\mathbf{X} = (X_{t_1,1}, \ldots, X_{t_1,j}, \ldots, X_{t_T,j})$, where $X_{t_i,j}$ is the jth replicate of pathogen genomes collected from wastewater at time t_i, with units of copies per milliliter. We model $X_{t_i,j}$ as a noisy representation of the unobserved number of currently infectious and recently recovered individuals. Let $\mathbf{O} = (O_{t_1}, O_{t_2}, O_{t_3}, \ldots, O_{t_T})$, where O_{t_u} is the number of newly observed cases of an infectious disease during time interval $(t_{u-1}, t_u]$. We model O_{t_u} as a noisy realization of the number of individuals transitioning from the latent stage of infection to the infectious stage.

4.2.2 Standard Compartmental Models

An SEIR compartmental model describes a homogeneously mixing population moving through infectious disease stages, referred to as compartments [Keeling and Rohani, 2008, pages 12-52]. The compartments are S, susceptible individuals; E, infected but not yet infectious individuals; I, currently infectious individuals; and R, no longer infectious either due to recovery or death. In its deterministic form, the changes in the number of individuals in these compartments are described using a system of ordinary differential equations (ODEs). The behavior of the ODEs is described by a set of key parameters defined in Table 4.1. The SEIR system of ODEs is:

$$\frac{dS}{dt} = -\beta_t \times I \times S/N, \quad \frac{dE}{dt} = \beta_t \times I \times S/N - \gamma \times E, \quad \frac{dI}{dt} = \gamma \times E - \nu \times I, \quad \frac{dR}{dt} = \nu \times I.$$

We model β_t as time-varying to account for changes in transmission due to, for example, implementation of public health policies, changes in behavior, or the emergence of new pathogen variants.

For the SEIR model, the time-varying basic reproduction number, $R_{0,t}$, the average number of individuals a person infected at time t would infect in a completely susceptible population, and effective reproduction number, R_t, are defined as:

$$R_{0,t} = \frac{\beta_t}{\nu},\ R_t = R_{0,t} \times \frac{S(t)}{N}. \tag{1}$$

We will adapt this classic model for our purpose of estimating R_t from wastewater data.

4.2.3 The EIRR Model

The SEIR model assumes that the susceptible population only changes because of new infections. In practice, the susceptible population can change over time due to vaccination campaigns and the introduction of new disease variants that evade immunity from prior infection. Modeling such dynamics is often difficult, especially in real time, when it may be non-trivial to estimate the effect of immune evasion of a particular variant.

Taking inspiration from birth-death modeling in infectious disease phylodynamics [Stadler et al., 2013], we define $\alpha_t = \beta_t \times S/N$ and rewrite the equation for dE/dt so that: $dE/dt = \alpha_t \times I - \gamma \times E$. The rate of new latent infections no longer depends on the S compartment. The parameter α_t can be interpreted as a time-varying average number of secondary infections produced by one infectious individual per unit time (e.g., per day). Note, the effective reproduction number is still recoverable, as $R_t = \beta_t S(t)/\nu N = \alpha_t/\nu$.

In addition, we split the R compartment in two. In the first compartment individuals are recovered but still shedding pathogen genomes, in the second they are recovered and no longer shedding genomes. This choice is motivated by the characteristics of SARS-CoV-2, for which it has been shown individuals shed detectable amounts of RNA in fecal matter well after the likely end of their infectious period (see Appendix Section B.2.3 [Okita et al., 2022, Zhang et al., 2021]). The final model, which we call the EIRR model, is described by the following equations:

$$\frac{dE}{dt} = \alpha_t \times I - \gamma \times E, \quad \frac{dI}{dt} = \gamma \times E - \nu \times I, \quad \frac{dR1}{dt} = \nu \times I - \eta \times R1, \quad \frac{dR2}{dt} = \eta \times R1. \tag{4.1}$$

Here $1/\eta$ is loosely interpreted as the average time spent recovered but still shedding pathogen RNA via fecal matter. We also include a redundant compartment $C(t)$, where $dC/dt = \gamma \times E$. This counts cumulative transitions from the E to I compartments, and allows us to keep track of the number of people who became infectious during time period $(t_u, t_{u-1}]$ as $C(t_u) - C(t_{u-1})$. For the sake of comparison, we will also implement the SEIRR model, which is the SEIR model with two R compartments.

4.2.4 Wastewater Observation Model

We model the log of observed pathogen genome concentrations as realizations of a generalized t-distribution: $\log X_{t_i,j} \sim \text{Generalized t}(\log(\lambda \times I(t_i) + (1-\lambda) \times R1(t_i)) + \log(\rho), \tau^2, df)$. Here $I(t_i)$ is the number of currently infectious individuals at time t_i, $R1(t_i)$ is the number of non-infectious but still shedding individuals at time t_i. The parameter $\lambda \in (0,1)$ is a normalized weight representing how much each individual contributes to the true underlying pathogen genome concentrations while infectious (Appendix Section B.1.1 describes in detail how we chose the prior for this parameter). Parameter ρ allows for flexibility in relating counts of individuals to concentrations. Parameter τ accounts for variation from the mean, and df is the parameter governing the degrees of freedom of the t-distribution. We chose to use a t-distribution because wastewater data often has many outliers; a t-distribution with thicker tails should better fit the data as opposed to the normal distribution.

4.2.5 Complete EIRR-ww Model Structure

We describe the complete wastewater data model, which we call the EIRR-ww model, structure in the following section. We use a random walk prior for the time-varying effective reproduction number: $R_0 \sim \text{Log-Normal}(\mu_0, \sigma_0), \sigma \sim \text{Log-Normal}(\mu_{rw}, \sigma_{rw}), \log(R_{k_i}) | R_{k_{i-1}}, \sigma \sim \text{Normal}(\log(R_{k_{i-1}}), \sigma)$. The times k_i can be chosen flexibly, for this chapter, we choose them so that R_t changes on a weekly basis. Let $\bar{\mathbf{\Theta}} = (\gamma, \nu, \eta, I(0), E(0), R1(0))$ and $\mathbf{R} = (R_{k_1}, \ldots, R_{k_M})$ be the vector of effective reproduction number values. Let $M(t, \bar{\mathbf{\Theta}}, \mathbf{R}) = (\mathbf{E}(t, \bar{\mathbf{\Theta}}, \mathbf{R}), \mathbf{I}(t, \bar{\mathbf{\Theta}}, \mathbf{R}), \mathbf{R1}(t, \bar{\mathbf{\Theta}}, \mathbf{R}), \mathbf{R2}(t, \bar{\mathbf{\Theta}}, \mathbf{R}))$ be the solution to the EIRR ODE system described in Section 4.2.3. The target posterior distribution is:

$$P(\mathbf{R}, \bar{\mathbf{\Theta}}, \rho, \lambda, \tau, df, \sigma \mid \mathbf{X}) \propto \underbrace{P(\mathbf{X} \mid \mathbf{M}(t, \bar{\mathbf{\Theta}}, \mathbf{R}), \rho, \lambda, \tau, df)}_{\text{Concentration Model}} \underbrace{P(\mathbf{R} \mid \sigma)}_{\text{RW Prior}} P(\bar{\mathbf{\Theta}}, \rho, \lambda, \tau, df, \sigma).$$

We use the No-U-Turn Sampler, implemented in the `Julia` package `Turing` to approximate this posterior distribution [Hoffman and Gelman, 2014, Ge et al., 2018]. We used non-centered re-parameterizations for all model parameters except for *df* (which had a gamma prior). Markov chain Monte Carlo chains were initialized using the Maximum A Posterior (MAP) estimate of each parameter plus Gaussian noise (except for *df* which was only initialized at the MAP).

4.2.6 The EIR Model

The EIRR model can be simplified when fitting to case data by using only a single R compartment, creating the EIR model. It is described by the same equations as equation 4.1 but with only one R compartment equation. Cases are modeled as a noisy realization of the number of transitions from the E to the I compartment using a negative-binomial likelihood. For cases observed in the interval $(t_{u-1}, t_u]$: $O_u \sim$ Negative-Binomial$((C(t_u) - C(t_{u-1})) \times \psi, \phi)$ where $C(t_u) - C(t_{u-1})$ is the number of transitions from the E to the I compartment in time interval $(t_{u-1}, t_u]$, ψ is a detection rate parameter, and ϕ is an over-dispersion parameter. Both ψ and ϕ have their own priors. The full structure of our case model, the EIR-cases model, is otherwise very similar to the EIRR-ww model. The structures of the corresponding SEIR-cases/SEIRR-ww models are likewise similar, though for SEIR-cases/SEIRR-ww models the basic reproduction number is modeled as random walk, rather than the effective reproduction number. We provide a more detailed description of the SEIRR-ww model priors in Appendix Section B.1.3.

All code used to produce this paper is available at `https://github.com/igoldsteinh/ww_paper`. A `Julia` package implementing the models used in this paper is available at `https://github.com/igoldsteinh/concRt.jl`. An `R` package which provides a wrapper for the Julia package is available at `https://github.com/igoldsteinh/concRt`.

4.3 Simulation

4.3.1 Simulation Protocols

We simulated a single realization from an agent-based stochastic SEIRR described in Appendix Sections B.1.1 and B.2.1. The population size was set to 100,000. The mean latent period was 4 days, the mean infectious period was 7 days, and the mean time spent recovered but still shedding pathogen genomes was 18 days. The simulation was started with 200 individuals in each of the E and I compartments, and run for a warm-up period of 77 days before creating data for the model to fit to. This was done so that there would be individuals in all compartments for whom all transition times between compartments would be naturally available. The top left panel of Figure B.8 shows the prevalence in each compartment, the time 0 is the day before the first wastewater sample is collected. The basic reproduction number $R_{0,t}$ was given a fixed trajectory. We calculated the true R_t at each day using Equation 1. For the observation period, we chose to start with $R_{0,t}$ set to 0.9 with a rapid increase to 2.5, where it stayed for the duration of the simulation, mimicking a scenario where a new and highly infectious variant is introduced into a population. All priors used in the simulation are listed in Table B.2 (for the case models the priors for γ and ν are transformed to be on a weekly scale). Note that the prior for λ was centered at 0.99, with a 95% quantile range of 0.8 to 1.

Using this single realization from the stochastic SEIRR model, we simulated 100 data sets of pathogen genome concentrations and 100 data sets of observed case data. All parameters specified below were chosen to create data similar to observed data from the SARS-CoV-2 pandemic in Los Angeles, California (see Appendix Section B.2.3 for more details). Daily genome concentration data were generated using a generalized t-distribution as in the model described in Section 4.2.4. However, the mean of the generalized t-distribution was the true total genome concentration shed, generated using the method described in Appendix Section B.1.1. We simulated ten replicates per day. Parameter ρ was set to be 0.011, τ was set to be 0.5, and df was set to 2.99. Only data from every other day were used for analysis. Cases were simulated at a daily time-scale and aggregated to a weekly time-scale. The case detection rate ψ was set to 0.2, while ϕ was set to 57.55. We generated data for a total of 19 weeks. While the case likelihood of our SEIR-cases and EIR-cases models is quite similar to the data generating mechanism of the simulated data, the wastewater likelihood of our SEIRR-ww and EIRR-ww models is a crude approximation of the data generating mechanism of the simulated data. For all subsequent simulation scenarios except the final one, we use the same 100 data sets while changing either the type of data used to fit the model, or the model priors.

In the baseline simulation scenario, we fit the SEIR-cases, EIR-cases, SEIRR-ww and EIRR-ww models to the data, using three replicates for the wastewater models and weekly cases for the case models. In subsequent scenarios, we only fit the EIRR-ww model. Using the same priors as in the baseline scenario, we fit the EIRR-ww model using one or ten replicates instead of three replicates (1-rep and 10-rep) and also fit the model using the mean of three replicates or the mean of ten replicates (3-mean and 10-mean respectively). We also conduct sensitivity analyses where the priors for the inital E and I compartments are centered at 75% or 133% of true values (Low Init and High Init respectively). We shift the prior for λ so that it is centered around 0.8 (Low Prop) as opposed to the default 0.99. We fit the Huisman method using the mean of three replicates as the input data. Details on choosing parameters for the Huisman method are in Appendix Section B.1.6. In the final scenario (Stoch R_t), we use the parameters, data, and priors of the baseline scenario, but for each data set simulate a new epidemic, and thus a new R_t curve, for each simulation. An example realization of the simulation is displayed in Figure B.8. All priors used in the baseline simulation are listed in Table B.2.

4.3.2 Comparison with State-of-the-Art Methods

We compare the EIRR-ww model to the Huisman et al. [2022b] method. This method is a variation on the well known `EpiEstim` method [Cori et al., 2013]. Pathogen concentrations are modeled as a convolution of unobserved latent incidence (new infections) and the individual shedding load profile describing how many gene copies individuals shed over the course of their infection. Latent incidence is estimated using an EM algorithm, then the estimated incidence is used as the input into `EpiEstim`. The pipeline is repeated multiple times using a bootstrap method to produce final measures of uncertainty. Further details are available in Appendix Section B.1.2.

In contrast, the method of Nourbakhsh et al. [2022] uses a compartmental model similar to ours, splitting both the I and $R1$ compartments into many smaller compartments in order to better match the shedding dynamics of pathogen genomes in fecal matter. They also directly model the impact of the sewer system itself on the final observed data. We decided not to test the method of Nourbakhsh et al. [2022] in this chapter because the code for the latter method is not readily available, and because the goal of their model was not limited to inference of the effective reproduction number.

4.3.3 Simulation Results

Baseline Simulation

Posterior medians and credible intervals for models fit to the example simulation data are displayed in Figure 4.1.

Posterior trajectories from all models generally mimic the true R_t curve, although the SEIRR-ww model struggles to capture the exact trajectory. Models using genome concentrations have wider credible intervals than models using case counts, reflecting higher variability of wastewater data as compared to case data. Note that the random walk prior forces R_t to change on a weekly scale, while the true values are reported on a daily scale, resulting in more pronounced segmentation of posterior summaries. The SEIRR-ww model (top left of Figure 4.1) estimates sloping spikes because the SEIRR ODEs are solved at a daily time scale, since the SEIR-cases model is solved at a weekly scale, it does not display the same behavior. EIRR-ww posterior estimates of the latent trajectories (including latent incidence) are displayed in Appendix Figure B.9. While posterior estimates mimic the shape of the latent trajectories, they fail to capture the magnitude of the trajectories except at the beginning and end of the simulation.

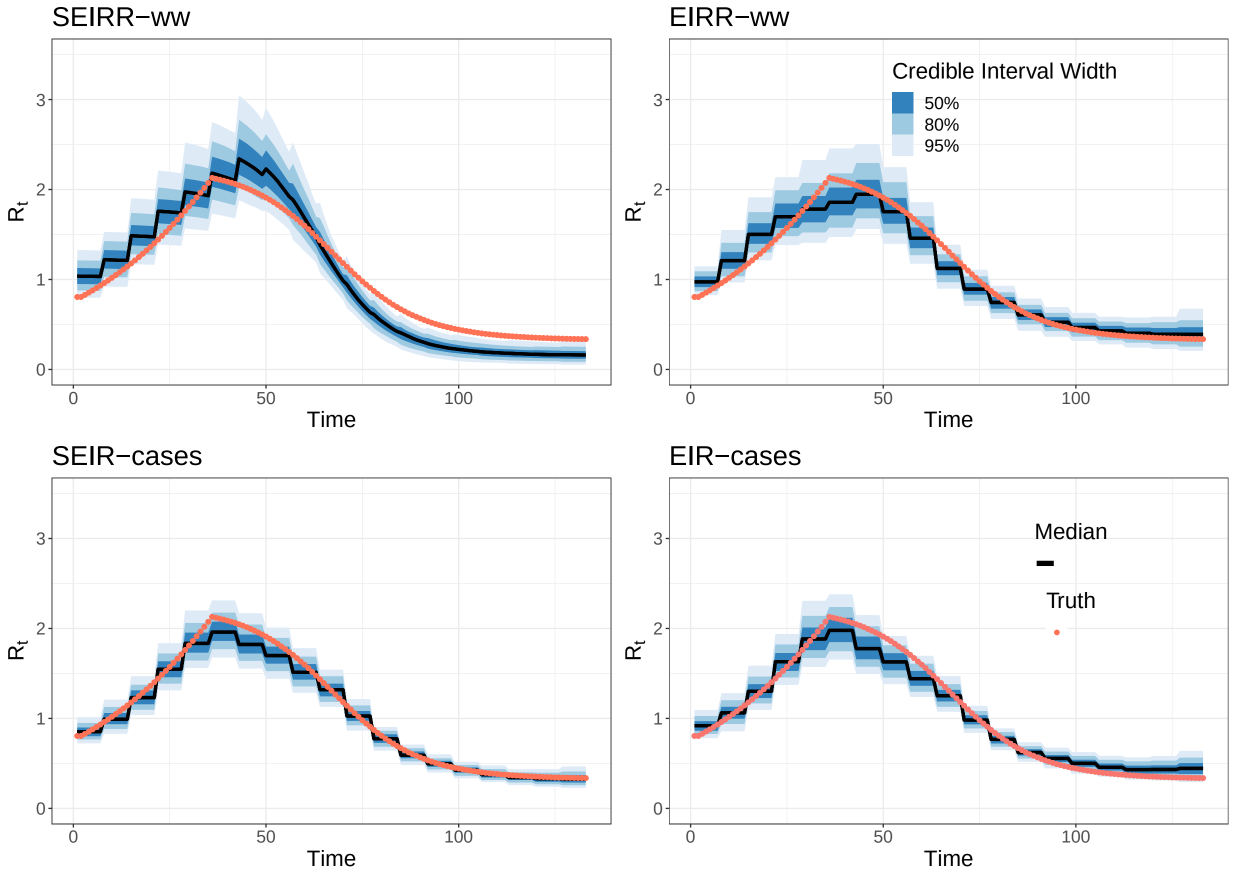


Figure 4.1: Posterior summaries of Rt using four models fit to either wastewater or case data generated from the same underlying infectious disease dynamics. True Rt trajectories are colored in red, black lines represent posterior medians, blue shaded areas from dark to light are 50, 80 and 95% credible intervals. Models in the top row use genome concentrations, models in the bottom row use case counts. Models in the left column use the S compartment, models in the right column do not.

To assess performance across many simulated data sets, we examined frequentist properties of our four models, summarized in Figure 4.2. Boxplot solid lines represent medians, hinges are upper and lower quartiles and whiskers are at most 1.5 times larger than the upper and lower quartiles. Envelope is a measure of coverage. For each simulation the envelope is the proportion of time points for which an 80% credible interval from the posterior distribution captured the true value of interest. Ideally it should be 0.8. We chose to use 80% credible intervals as estimates of the 80% quantiles have less Monte Carlo Error than 95% quantiles, so fewer data sets are needed to estimate them well. The corresponding 95% credible interval results are displayed in Appendix Figure B.13. Mean credible interval width (MCIW) is the mean of 80% credible interval widths across time points within a simulation. Absolute deviation is a measure of bias, and is the mean of the absolute difference between the posterior median and the true value at each time point. Finally, mean absolute sequential variation (MASV) measures how well each method captures the variation in the effective reproduction number across time by computing the mean of the absolute difference between the posterior median at t and the posterior median at $t-1$. We compare this to the true mean absolute sequential variation in each simulation. The EIRR-ww model outperforms the SEIRR-ww model in terms of bias, precision, and coverage. Both the SEIR-cases model and the EIR-cases model outperform the EIRR-ww model in terms of bias and precision. The EIR-cases model is slightly more biased and less precise than the SEIR-cases model. Although the EIRR-ww model has less precision than case based models, this simulation shows that models using wastewater data can be used to estimate the effective reproduction number reasonably well.

Performance in Other Scenarios

Frequentist metrics comparing the baseline EIRR-ww fit (three replicates) to the EIRR-ww fit to one and ten replicates, and to the EIRR-ww fit to mean of three replicates and mean

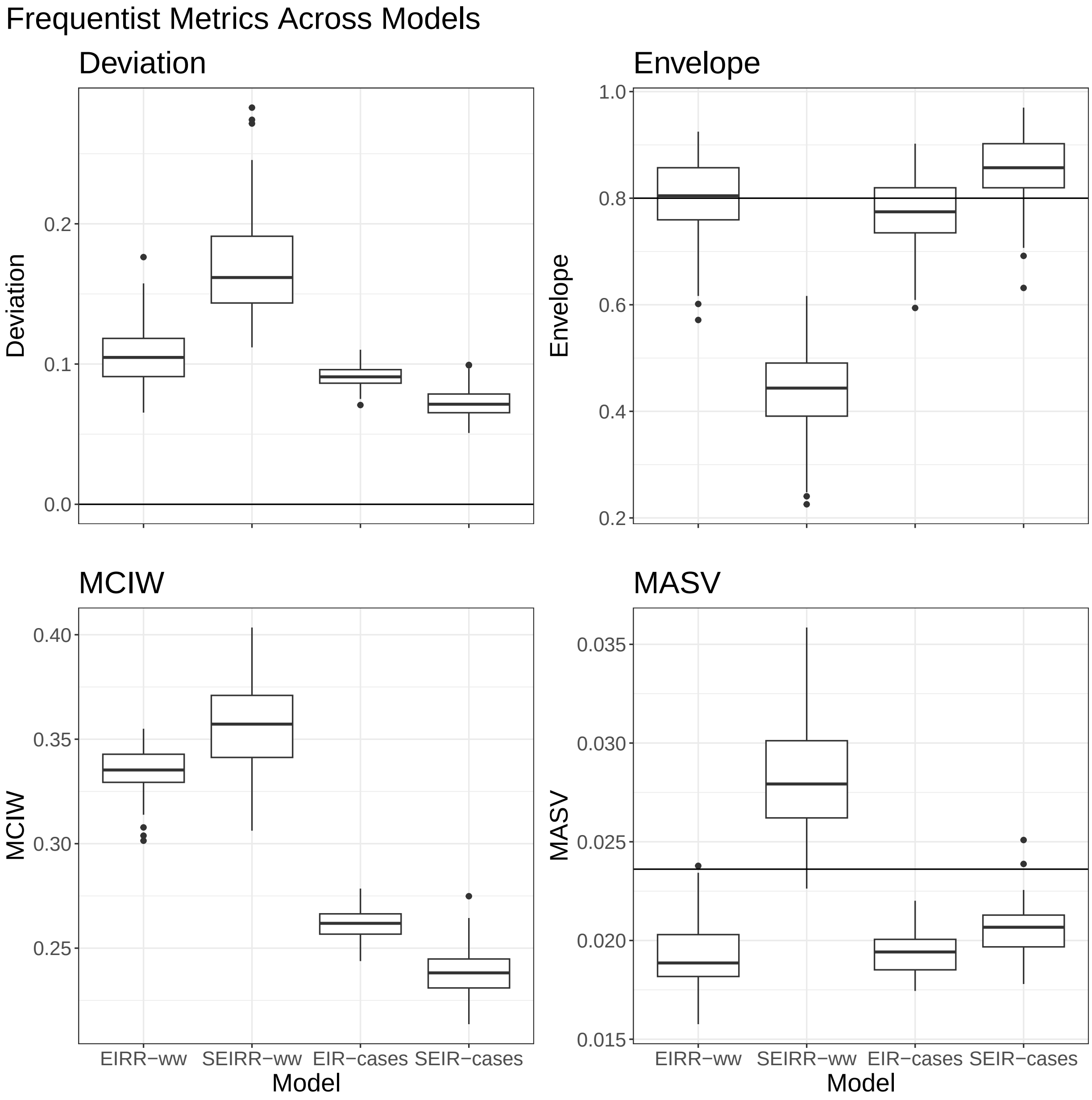


Figure 4.2: Frequentist metrics for the EIRR-ww, SEIRR-ww, EIR-cases and SEIR-cases models in the baseline scenario. Absolute deviation is the mean of the absolute value of the difference between the median R_t at each time point and the true value. Envelope is a measure of coverage, taking the average coverage of 80% intervals over the time series. MCIW is the average mean credible interval width. Mean absolute standard deviation (MASV) is the difference between the current median point estimate for R_t and the previous point estimate for R_t. The line in the bottom right panel represents the true absolute standard deviation. Solid lines represent medians, hinges are upper and lower quartiles and whiskers are at most 1.5 times the inter-quartile range from the median.

of ten replicates, are displayed in Appendix Figure B.14. The EIRR-ww fit to one or ten replicates performed modestly worse or better respectively than the EIRR-ww fit to three replicates. The EIRR-ww fit to raw replicate concentrations performed slightly better than the EIRR-ww fit to means of replicates.

We assessed the robustness of our model by changing the priors for the initial conditions, as well as the prior for λ. Frequentist metrics comparing these alternate models to the baseline model are displayed in Figure B.15. Changing these priors lead to only modest changes in model performance.

We fit the method by Huisman et al. [2022b] to each of our data sets, using as the input data the mean of three replicates. We then compared the Huisman et al. [2022b] method to the EIRR-ww model fit to three replicates. Also, the Huisman et al. method does not provide 80% credible intervals, so we compared metrics using 95% credible intervals. The comparison is visualized in Figure B.16. In this simulation, the EIRR-ww model clearly outperforms the Huisman model in terms of both bias and precision associated with the effective reproduction number trajectory estimation. We found that the EIRR-ww model performed similarly to the baseline scenario when fit to 100 data sets where each data set was generated from a separate simulated epidemic and separate simulated R_t. The comparison is visualized in Figure B.17. Summaries of MCMC diagnostics for all model fits are available in Appendix Section B.2.8.

4.4 The Effective Reproduction Number of SARS-CoV-2 in Los Angeles, CA

4.4.1 Data

Wastewater data were collected from the Joint Water Pollution Control Plant (JWPCP), one of the largest wastewater treatment plants in Los Angeles County. The plant serves 4.8 million people across Los Angeles County. The data, reported as viral gene copies per ml of wastewater determined by quantitative PCR, were collected from the 24-hour composite wastewater influent samples at irregular, but approximately two-day intervals, and usually three replicates were reported for each sample [Song et al., 2021]. We excluded two days (7/7/21 and 8/23/21) as outliers, as the reported concentrations dropped by at least two orders of magnitude compared to the concentrations of the closest previous and subsequent observed days. Cases during the same period in Los Angeles County are available from the California Open Data Portal [California Open Data Portal, 2023]. The available data from cases and wastewater are visualized in Figure 4.3. The cases are recorded for all of Los Angeles County, not just the population served by the JWPCP plant. We re-scaled the cases by a factor of 0.48 to partially account for this, as there are about 10 million people in Los Angeles County in total. Priors for the EIRR-ww and EIR-cases model were the same as those used in the simulation, except for the priors on the initial conditions and initial R_t (see Appendix Section B.2.10). For the Huisman et al. [2022b] method, we used the same shedding load profile calculated for the simulated data sets, but used the mean and and standard deviation of the generation time distribution of SARS-CoV-2 calculated by Sender et al. [2021].

4.4.2 Results

The posterior estimate of R_t (from left to right) of the Huisman et al. [2022b] method, the EIR-cases model and the EIRR-ww model are shown in Figure 4.4. For additional comparisons, we used two branching process models, the Rt-estim-gamma model fit to cases and total number of diagnostic tests (Chapter 3) and another model fit to cases using the `epidemia` package [Bhatt et al., 2023, Scott et al., 2021]. For more details on these models, see Appendix Section B.1.7. Posteriors from these models, along with the EIR-cases and EIRR-ww models are shown in in Appendix Figure B.18.

While the EIRR-ww model provides different estimates than any of the case-based methods, they mostly align, estimating one large increase in R_t above 1 tied with the arrival of the Omicron variant in California in winter 2021. In contrast, the Huisman method estimates several dramatic changes in R_t over short spans of time, most notably an increase from below 1 to above 2 at the start of October 2021. Overall, we think the EIRR-ww model provides a better estimate of R_t than the Huisman method.

In addition, we calculate the case detection rate normalized by total diagnostic tests to account for changes in the case detection rate due to changes in available diagnostic tests. The posteriors are visualized in Figure B.19. Both normalized and un-normalized versions of the estimated posterior case detection rate changed dramatically during the observation period. More details are available in Appendix Section B.2.12.

4.5 Discussion

We have presented a modeling framework for using simplified compartmental models coupled with Bayesian non-parametric priors to estimate the effective reproduction number. Using this framework, we created the EIRR-ww model to estimate the effective reproduction num-

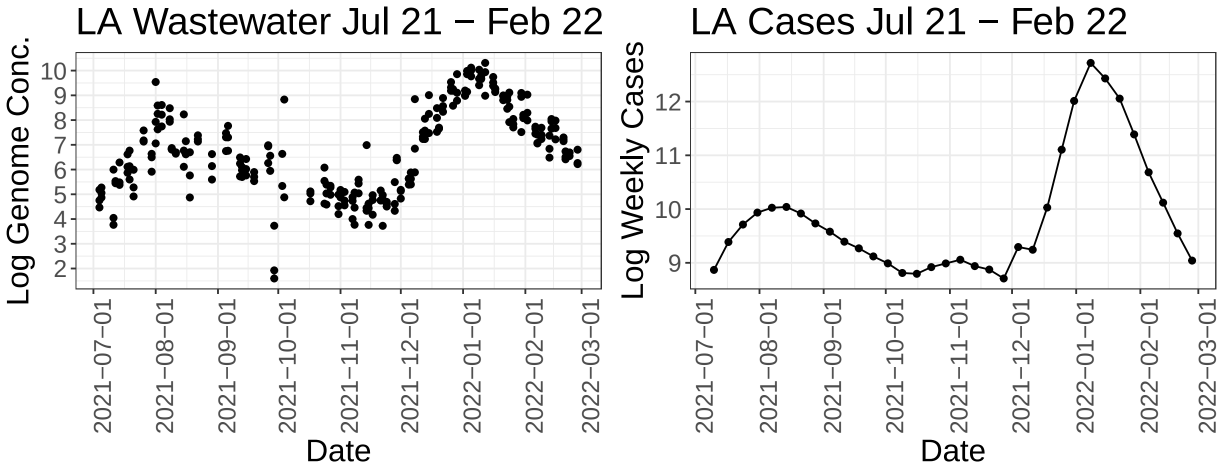


Figure 4.3: Wastewater and case data for the SARS-CoV-2 epidemic in Los Angeles County. Wastewater data were collected approximately every two days, with usually three measurements taken per day, each dot represents a measurement, the line is the average of the measurements. Cases were aggregated to a weekly time scale.

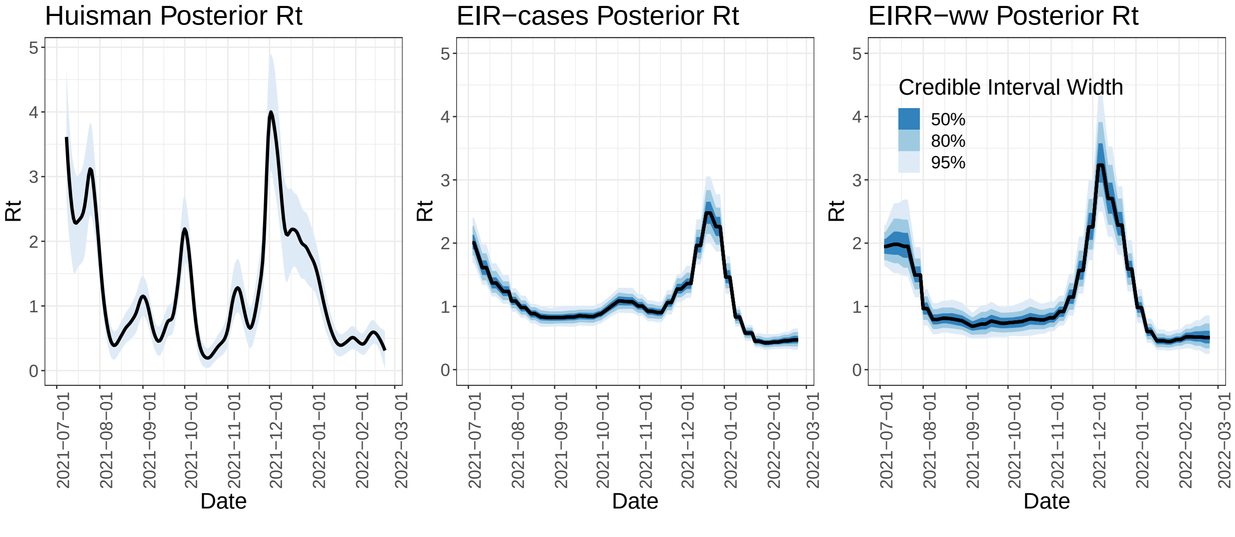


Figure 4.4: Posterior summaries of the effective reproduction number for the SARS-CoV-2 Delta and Omicron waves in Los Angeles, CA. Blue bars from dark to light represent 50, 80, and 95% credible intervals. Black lines represent median posterior estimates. EIRR-ww model and Huisman model are fit to RNA concentrations collected from wastewater data, EIR-cases model is fit to weekly case counts.

ber using pathogen genome concentrations collected from wastewater samples. We tested the EIRR-ww model by fitting it to data simulated from an agent-based stochastic SEIRR model, and showed it could successfully estimate effective reproduction number dynamics. We also used the EIRR-ww model to estimate the effective reproduction number for SARS-CoV-2 in Los Angeles, California, showing it provides plausible estimates of the effective reproduction number when used on real world data.

Our proposed models ignore the individual time-varying shedding dynamics of pathogen genome concentrations. The SEIRR-ww model struggled to estimate the effective reproduction number on simulated data that accounted for the time-varying nature of the data. In contrast, the EIRR-ww model had smaller bias and credible interval widths and was well calibrated from a frequentist perspective. The gap in performance likely stems from the fact that both models use mis-specified observation models and the EIRR-ww model's lack of an S compartment results in more flexibility for its estimates of the effective reproduction number, and thus better performance overall. However, the EIRR-ww model was unable to reconstruct the latent population compartments using only the linear relationship between compartment counts and concentrations (Figure B.9).

Our method clearly outperformed the Huisman et al. [2022b] method on both simulated and real data. We speculate the high levels of noise in both simulated and real data sets resulted in overly jagged estimates from `EpiEstim`, a problem we demonstrate directly in Chapter 3. Our method is an appealing alternative when the main goal is effective reproduction number inference.

When compared to models fit to case data, the EIRR-ww model clearly had larger bias and wider credible intervals, likely due to the high individual variation in genome concentrations seen in empirical studies [Hoffmann and Alsing, 2023]. On the other hand, in real world settings where the case detection ratio changes over time, models fit to wastewater data may outperform models fit to case data that do not account for this. Using models fit to both cases and wastewater data simultaneously is a promising direction for future work.

The EIR-cases model had slightly larger bias and wider credible intervals than the SEIR-cases model. However, this comparison was made without considering waning immunity. We did not use the SEIR-cases model to estimate R_t for Los Angeles, CA, in part because we think it is highly likely the rates of waning immunity changed dramatically when Omicron became the dominant SARS-CoV-2 variant. This is a situation for which the EIR framework is well suited. Even without wastewater data, we think the framework we describe in this paper is a useful alternative to `EpiEstim` and related methods when estimating the effective reproduction number.

We used both the EIRR-ww model as well as three other case based models to estimate the effective reproduction number of SARS-CoV-2 in Los Angeles, CA. We emphasize that all methods have their limitations, and none should be taken as ground truth. However, the models agreed at many key points in time, including when R_t falls below one after the summer 2021 wave, and the general timing of the winter 2021 wave. This agreement suggests the EIRR-ww model estimates are not unreasonable when fit to real data. For a fuller discussion of the points of disagreements between the models, see Appendix Section B.3.1.

In this chapter we focused on replicates, rather than the commonly reported averages of replicates and found that, when using simulated data, models using replicates performed modestly better than models relying on average concentrations [Duvallet et al., 2022, Wastewater-SCAN, 2023]. While the improvements in performance are not large, these improvements are basically free, as the data are already being collected. We also found that using ten replicates instead of three produced only modestly improved model performance. Depending on the cost of producing replicates, increasing the number of replicates sampled may not be worth pursuing.

For SARS-CoV-2, most shedding occurs in the infectious period, raising the possibility our model could be simplified to exclude the R1 compartment. This would simplify our closed form solutions and speed computations, and is a promising modification of the model to explore further. However, including multiple compartments may be necessary for other pathogens with different shedding profiles. We chose not to incorporate covariates that can control for changes in population size and conditions in the sewer system into our model. For our particular application, the JWPCP plant is so large, and collects wastewater from so many different smaller plants, that the population size and conditions are plausibly stable across time. In addition, there remains some controversy over exactly which covariates would be most useful to include [Maal-Bared et al., 2023]. Incorporating covariates which adjust for these changes is an important next step.

Chapter 5

The Signal is Not Flushed Away: Inferring the Effective Reproduction Number From Wastewater Data in Small Populations

5.1 Introduction

Pathogen genome concentrations collected from wastewater (henceforth referred to as wastewater data) provide information about the number of currently infected and recently covered individuals in an infectious disease outbreak, and thus are a potential source of data when modeling an epidemic [Hillary et al., 2020, Polo et al., 2020]. While there are many examples of studies that correlate wastewater data with other data sources, such as Song et al. [2021], there are relatively few instances of wastewater data being incorporated into epidemiological statistical models ([Huisman et al., 2022b, Nourbakhsh et al., 2022, Morvan et al., 2022] and

Chapter 4). To our knowledge, what studies do exist have largely focused on large populations, harvesting data from large wastewater treatment plants that collect wastewater from hundreds of thousands of individuals. However, epidemiological modeling of small populations is also of interest, and while there has been interest in using wastewater data to study infectious disease outbreaks in small populations, such as long term care facilities or college dormitories [Keck et al., 2024, Acer et al., 2022], to our knowledge, there are no statistical methodologies that address this challenge. In this paper, we provide a novel method for estimating the effective reproduction number in small populations from wastewater data.

Ideally, we would model the spread of infectious diseases as a Markov Jump Process (MJP), dividing a population into compartments representing susceptible, infectious and recovered individuals, but in practice this is computationally challenging when the state space is large and the process only partially observed [Ho et al., 2018, Rupp et al., 2024]. In large populations, deterministic approximations, such as using a model described by ordinary differential equations (ODEs) can be used, but in small populations, these deterministic approximations can be inadequate. Previous methods for estimating the effective reproduction number from wastewater data have relied on these deterministic approximations, and thus may perform poorly in small population settings (Chapter 4 and [Huisman et al., 2022b]).

Exact inference of MJPs for epidemics is usually accomplished either through Sequential Monte Carlo or data augmentation [O'Neill and Roberts, 1999, King et al., 2016, Andrieu et al., 2010]. Both approaches can be computationally intensive and prone to implementation issues in practice, and improving these methodologies remains an active area of research [Corbella et al., 2022a, Morsomme and Xu, 2022]. It is also common practice to approximate the MJP with the linear noise approximation, a local approximation to a stochastic differential equation whose solution is itself a diffusion process approximation of the MJP [Fintzi et al., 2022, Golightly et al., 2023]. The key advantage of the linear noise approximation is that it has Gaussian transition densities. Another approach, proposed by Isham [1991], is to start with the Gaussian transition densities (justified under central limit theorem style arguments originating from Kurtz [Kurtz, 1971, Britton and Pardoux, 2019]), and then choose approximate first and second moments of the MJP to use as the moments of the Gaussian density [Isham, 1991, Buckingham-Jeffery et al., 2018].

For the specific task of inferring the effective reproduction number, it is quite common to choose an approximate model that does not model the number of susceptible individuals explicitly. The most common methods are based on a branching process approximation to the MJP where individuals infect other individuals in an independent and identically distributed manner, giving rise to the commonly used renewal equation relating the current number of new cases to the previous counts of new cases ([Cori et al., 2013, Bhatt et al., 2023] and Chapter 3). Avoiding modeling susceptibles simplifies the model significantly, while still allowing for accurate inference of the effective reproduction number. Because these branching process based models are based on incidence, it can be cumbersome to connect them to data sources that are not explicitly realizations of incidence, such as wastewater data. We previously developed a compartmental model that did not model susceptibles, but our approach used a deterministic model which may not be appropriate for small populations (Chapter 4).

We adapt our previous approach by first defining an MJP without a susceptible compartment. This in turn greatly simplifies the use of either the LNA or the moment approximation technique, as the infinitesimal transition rates are now linear. In this paper, we chose to use the approach of Isham [1991], taking advantage of the simplified form of the MJP to calculate the exact conditional first and second moments in closed form. Our approach allows us to make use of state-of-the-art high dimensional Markov chain Monte Carlo methods that scale well with population size while still accounting for the stochastic nature of infectious disease transmission.

We compare our stochastic model to its deterministic counterpart under multiple simulation scenarios. We first vary the shape of the curve we are trying to infer, then change the "stochasticity" of the underlying epidemic dynamics by varying the initial number of infected individuals and the total population size. Finally, we apply our wastewater based methods to estimate the effective reproduction number of SARS-CoV-2 in several college campus dormitories and compare the results with state-of-the-art case-based methods, demonstrating the qualitative differences between wastewater and case based methods, as well as the differences between deterministic and stochastic methods in a small population setting.

5.2 Methods

5.2.1 Wastewater Data

It is common practice to measure the concentration from the same sample of wastewater multiple times, producing multiple measurements called replicates. In real world data sets, an average of replicates is often reported, but we will instead focus on the replicates, as there are some advantages to using the raw data as opposed to an average (Chapter 4). We define $\mathbf{X} = (X_{t_1,1}, \ldots, X_{t_1,j}, \ldots, X_{t_T,j})$, where $X_{t_i,j}$ is the jth replicate of pathogen genomes collected from wastewater at time t_i, with units of copies per milliliter. We will model $X_{t_i,j}$ as a noisy realization of the unobserved number of currently infectious individuals.

5.2.2 Stochastic Compartmental Models

5.2.3 The SEIR Model

The SEIR model describes an infectious disease outbreak of a homogeneously mixing population, with the population divided into four compartments: susceptible, exposed (infected but not yet infectious), infectious, and removed. In its stochastic form, we represent the SEIR model as a four dimensional continuous time Markov jump process, $\mathbf{G(t)} = (S(t), E(t), I(t), R(t))$. It can be defined in terms of rate parameters such that

$$
\begin{aligned}
&P(\mathbf{G}(t+dt) = (s-1, e+1, i, r) \mid \mathbf{G}(t) = (s, e, i, r)) = \beta \times i \times s/N \times dt + o(dt),\\
&P(\mathbf{G}(t+dt) = (s, e-1, i+1, r) \mid \mathbf{G}(t) = (s, e, i, r)) = \gamma \times e \times dt + o(dt),\\
&P(\mathbf{G}(t+dt) = (s, e, i-1, r+1) \mid \mathbf{G}(t) = (s, e, i, r)) = \nu \times i \times dt + o(dt)\\
&P(\mathbf{G}(t+dt) = (s, e, i, r) \mid \mathbf{G}(t) = (s, e, i, r)) = 1 - (\beta \times i \times s/N + \gamma \times e + \nu \times i)dt + o(dt).
\end{aligned}
$$

Here γ is the inverse of the mean latent period, and ν is the inverse of the mean infectious period. We describe the infectiousness of the disease through the rate parameter β. In practice, we will allow β to be time-varying, and denote it β_t, to allow for changes in population or pathogen characteristics such as public health policies or emergence of more transmissible genetic variants. With this model, the time-varying basic reproduction number, $R_{0,t}$, and effective reproduction number, R_t, are defined as

$$
\begin{aligned}
R_{0,t} &= \frac{\beta_t}{\nu},\\
R_t &= R_{0,t} \times \frac{S(t)}{N}.
\end{aligned}
$$

5.2.4 The EI Model

We reduce the SEIR model to the EI model in order to avoid modeling changes in the S compartment due to changing immunity profiles caused by vaccines or new immunity evading variants. We represent the EI model as a two dimensional continuous time Markov jump process $\mathbf{H}(t) = (E(t), I(t))$, defined with rates as:

$$
\begin{aligned}
&P(\mathbf{H}(t+dt) = (e+1, i) \mid \mathbf{H}(t) = (e, i)) = \alpha_t \times i \times dt + o(dt),\\
&P(\mathbf{H}(t+dt) = (e-1, i+1) \mid \mathbf{H}(t) = (e, i)) = \gamma \times e \times dt + o(dt),\\
&P(\mathbf{H}(t+dt) = (e, i-1) \mid \mathbf{H}(t) = (e, i)) = \nu \times i \times dt + o(dt)\\
&P(\mathbf{H}(t+dt) = (e, i) \mid \mathbf{H}(t) = (e, i)) = 1 - (\alpha_t \times i + \gamma \times e + \nu \times i)dt + o(dt).
\end{aligned}
$$

Note that R_t is still recoverable by setting $\alpha_t = \beta_t \times \frac{S(t)}{N}$, so that

$$
R_t = R_{0,t} \times \frac{S(t)}{N} = \frac{\alpha_t}{\nu}.
$$

Ideally, we would like to have the transition probabilities of $\mathbf{H}(t)$ in closed form, but this is not analytically tractable. However, if we assume that α_t is piece-wise constant, then for any particular interval of time, the conditional moments of $\mathbf{H}(t)$ are available in closed form. We will use the conditional moments to construct approximations to the transition probabilities of $\mathbf{H}(t)$, following the techniques of Isham [1991].

5.2.5 Constructing a Partial Differential Equation of the Moment Generating Function

Let $p_{e,i} = P(\mathbf{H}(t) = (e,i) | \mathbf{H}(l) = (x, y))$ for $l < t$, we will omit indexing by x, y for notational simplicity. Then the Kolmogorov Forward equation for $\mathbf{H}(t)$ is

$$
\frac{dp_{e,i}}{dt} = \alpha I p_{e-1,i} + \gamma(E+1)p_{e+1,i-1} + \nu(I+1)p_{e,i+1} - (\alpha I + \gamma E + \nu I)p_{e,i}. \tag{5.1}
$$

From this differential equation with respect to time, we can construct a partial differential equation of the moment generating function of $\mathbf{H}(t)$, by multiplying both sides of Equation

5.1 by $e^{\theta_1 E+\theta_2 I}$, where θ_1, θ_2 take values in a subset of the real line which includes 0, and sum over all possible values of each of the four compartments. Let $M(\bar{\theta}; t)$ be the Moment-generating function of $\mathbf{H}(t)$. We produce the following partial differential equation:

$$\frac{dM(\bar{\theta};t)}{dt} = \left(\alpha e^{\theta_1}\frac{d}{d\theta_2} + \nu e^{-\theta_2}\frac{d}{d\theta_2} + \gamma e^{-\theta_1+\theta_2} + \frac{d}{d\theta_1} - \alpha\frac{d}{d\theta_2} - \nu\frac{d}{d\theta_2} - \gamma\frac{d}{d\theta_1}\right) M(\bar{\theta};t). \quad (5.2)$$

By taking the partial derivative on both sides, and utilizing properties of moment generating functions, we can create a system of linear ordinary differential equations for the conditional moments of the EI model (see the Appendix Section C.1 for the series of ODEs). We used Mathematica version 13.1 [Inc.] to generate closed form solutions of the conditional expectations, variances, and covariance. Let $\bar{\mu}$ be the vector of conditional expectations, and $\mathbf{\Sigma}$ be the matrix of conditional variances and covariances. We then use the derived conditional moments to construct densities that approximate the transition probabilities of the continuous time Markov jump process.

5.2.6 Log-Normal Approximation of the Transition Probabilities

We start with the known result that when the compartment counts are large enough, the transition probability mass function converges to the normal density, that is, for $l < t$

$$\mathbf{H}(t)|\mathbf{H}(l) \sim \text{Normal}(\bar{\mu}, \mathbf{\Sigma}). \quad (5.3)$$

In practice, we wish to rule out the possibility of negative compartment counts. To do this, we instead use the transition density for the log compartment counts, and appeal to the delta method to construct the density. That is:

$$\log \mathbf{H}(t)|\mathbf{H}(l) \sim \text{Normal}(\log \bar{\mu}, \nabla \log \bar{\mu} \mathbf{\Sigma} \nabla \log \bar{\mu}), \quad (5.4)$$

where $\nabla \log \bar{\mu}$ is the Jacobian of $\log \bar{\mu}$. We compare our Log-Normal approximate model against the true MJP empirically via simulation in Appendix Section C.2.

We place additional explicit priors on the latent state space for computational reasons. First, we require the compartment counts to be non-zero to avoid taking the log of 0. Second, we require the compartment counts and the means of the compartment counts to sum to less than 8 billion in order to avoid computations with infinity.

5.2.7 Observation Model

Recall that $X_{t_i,j}$ is the jth replicate of the concentration observed at time t_i. On the log scale, we model $X_{t_i,j}$ as a noisy realization of the number of currently infectious individuals, where

$$\log X_{t_i,j} \sim \text{Normal}(I(t_i) \times \rho, \tau^2). \tag{5.5}$$

Here ρ is a scaling factor, τ is a noise parameter, both receive priors.

Note that we are using a simpler model for pathogen concentrations than used in Chapter 4. We feel much of the benefit of using the t-distribution is lost when we have a stochastic latent epidemic process, and that, while our chosen model may not be ideal, the simplicity allows us to explore other aspects of the model well. Further discussion on this choice is in Appendix Section C.3.

5.2.8 Complete Stochastic EI-ww Model Structure

We use a random walk prior for the time-varying effective reproduction number: $R_0 \sim$ Log-Normal$(\mu_0, \sigma_0), \sigma \sim$ Log-Normal$(\mu_{rw}, \sigma_{rw}), \log(R_{k_i}) | R_{k_{i-1}}, \sigma \sim$ Normal$(\log(R_{k_{i-1}}), \sigma)$.

Let $\bar{\boldsymbol{\Theta}} = (\gamma, \nu, I(0), E(0))$, $\mathbf{R} = (R_{k_1}, \dots, R_{k_M})$ be the vector of effective reproduction number values and $\mathbf{H} = (\mathbf{H}_{t_i}, \dots, \mathbf{H}_{t_I})$ be the matrix of latent states from $\mathbf{H}(t)$. Note that we are interested in the states of $\mathbf{H}$ that correspond to our observed data, but we will also augment these states with additional states for any time at which α_t changes. The target posterior distribution is:

$$P(\mathbf{R}, \mathbf{H}, \bar{\boldsymbol{\Theta}}, \rho, \tau, \sigma \mid \mathbf{X}) \propto \underbrace{P(\mathbf{X} \mid \mathbf{H}, \mathbf{R}, \bar{\boldsymbol{\Theta}}, \rho, \tau)}_{\text{Concentration Model}} \underbrace{P(\mathbf{H} \mid \mathbf{R}, \bar{\boldsymbol{\Theta}})}_{\text{LN EI}} \underbrace{P(\mathbf{R} \mid \sigma)}_{\text{RW Prior}} P(\bar{\boldsymbol{\Theta}}, \rho, \tau, \sigma).$$

We use the No-U-Turn Sampler, implemented in the `Julia` package `Turing` to approximate this posterior distribution [Hoffman and Gelman, 2014, Ge et al., 2018]. We used non-centered re-parameterizations for all model parameters. Markov chain Monte Carlo chains were initialized using the Maximum A Posterior (MAP) estimate of each parameter with added independent Gaussian noise.

5.2.9 Other Models

We compare our stochastic EI-ww model to its deterministic counterpart. In the deterministic model, the compartments are modeled with a set of ODEs as follows

$$\begin{aligned} \frac{dE}{dt} &= \alpha I - \gamma E \\ \frac{dI}{dt} &= \gamma E - \nu I. \end{aligned}$$

Let $\mathbf{M}(\bar{\boldsymbol{\Theta}}, t)$ be the solution to the ODEs. Then the posterior of interest for the deterministic EI-ww model is

$$P(\mathbf{R}, \bar{\boldsymbol{\Theta}}, \rho, \tau, \sigma \mid \mathbf{X}) \propto P(\mathbf{X} \mid \mathbf{M}(\bar{\boldsymbol{\Theta}}, t), \mathbf{R}, \bar{\boldsymbol{\Theta}}, \rho, \tau) P(\mathbf{R} \mid \sigma) P(\bar{\boldsymbol{\Theta}}, \rho, \tau, \sigma).$$

There is no additional term for the compartment counts, because they are now deterministic.

In addition, when analyzing real data, we analyze case data with a model constructed using the `Epidemia` package Bhatt et al. [2023]. The model is an example of the common branching process based methods for estimating R_t from cases, where the mean number of new infections is equal to a weighted sum of the previous new infections multiplied by the effective reproduction number. The full model is written in Appendix Section C.4, and more thorough descriptions of this class of models can be found in Bhatt et al. [2023] and Chapter 3.

5.3 Simulations

5.3.1 Simulation Protocol

We simulated data from an agent-based stochastic SEIRR model that models each individual in a population, but is equivalent to a population level SEIRR model when aggregated (Appendix Section C.5.1). The additional extra R compartment allows for recently recovered individuals to shed pathogen RNA, a plausible characteristic of individuals infected with SARS-CoV-2 (Appendix Section B.2.3). The rates governing time spent in each compartment were chosen to mimic SARS-CoV-2 (Appendix Section B.2.3). Wastewater concentrations were generated using the normal distribution as in Section 5.2.7, however the mean of the distribution was the total genome concentration of the population, calculated by aggregating the individual concentration shed each individual in the population, which was allowed to vary over time to allow for the time-varying and individual heterogeneity of pathogen shedding observed in real world studies of SARS-CoV-2 (Appendix Section B.2.3).

For the first three simulations, the population size was 1000 and we simulated epidemics under three different R_0 curves, one where $R_{0,t}$ changed from 0.9 to 2.5 in five weeks (Steep), another where $R_{0,t}$ changed from 1 to 1.8 in seven weeks (Shallow), and finally one where

$R_{0,t}$ stayed fixed at 1.4 (Fixed). Note that while $R_{0,t}$ was fixed, because R_t is a function of the number of susceptibles $S(t)$, each simulation for a particular scenario had a different true R_t curve. We initialized ten individuals in the E and I compartments each, with the remainder in the S compartment. For the Fixed $R_{0,t}$ scenario, we simulated an additional four scenarios; one where the initial number of individuals in the E and I compartments was 5 or 20 individuals (Init5 and Init20 respectively), and finally two scenarios where the initial counts were 10, but the total population size was 500 or 2000 (Total500 and Total2000 respectively). These last four scenarios were meant to test how the models performed under varying levels of stochasticity. We would expect that smaller initial counts or total populations would have more stochastic variation, while the opposite would be true for larger counts. Models were fit to the first fourteen weeks of data. We used three replicates per day as data. For all scenarios, we model R_t as changing on a weekly basis. For each scenario, we simulated 100 epidemics. MCMC algorithms were initially run 800 or 1000 iterations for the stochastic and deterministic models respectively. If chains failed to pass basic convergence diagnostics, the models were re-fit for longer iterations. Note that for some simulations, the stochastic EI-ww model still failed basic convergence diagnostics: in particular four simulations for the Shallow scenario, three simulations for the Fixed scenario, three simulations for the Init5 scenario and one simulation for the Total500 scenario. Thes simulations were excluded in the final analysis, leaving ninety six, ninety seven, ninety seven and ninety nine simulations for analysis respectively. An example simulation from the Shallow curve setting is shown in Figure 5.1.

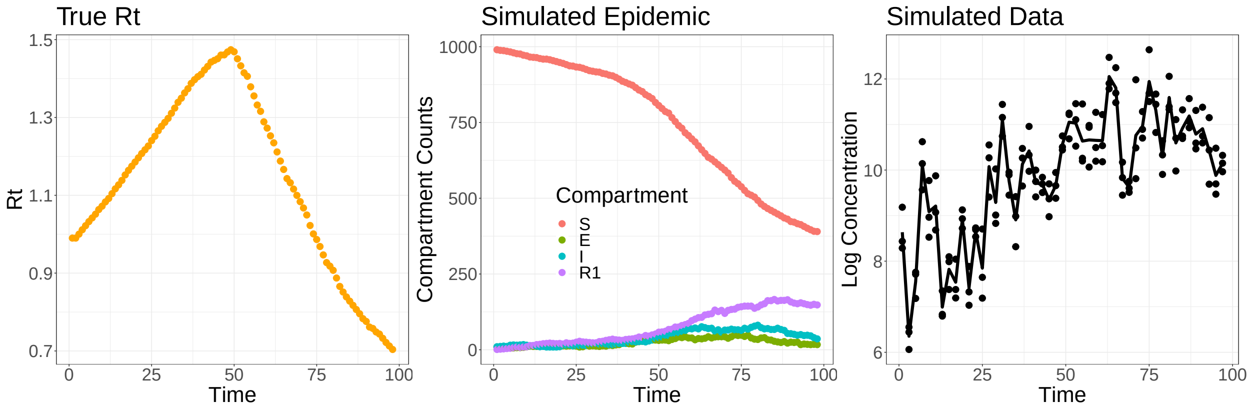


Figure 5.1: Simulated epidemic and corresponding wastewater data for the Shallow Rt scenario. The left panel displays the R_t, the middle panel shows the counts of individuals in each stage of infection, and the right panel shows the wastewater data simulated from this epidemic. In the third panel, the dots are replicates, while the black line represents the mean of the three replicates.

5.3.2 Simulation Results

Example posterior medians and credible intervals from the deterministic and stochastic EI-ww models for the Steep, Shallow, and Fixed scenarios are shown in Figure 5.2.

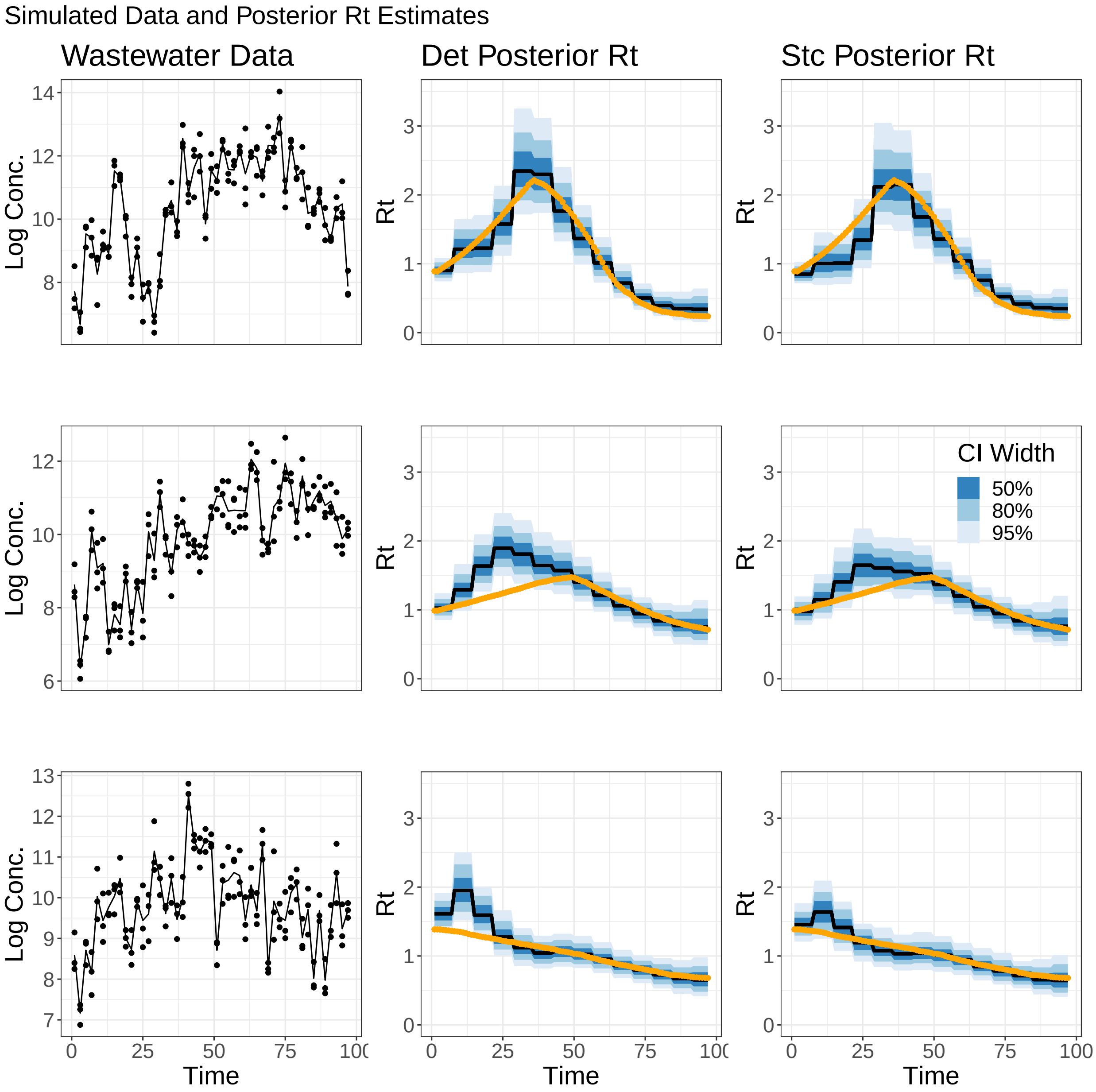


Figure 5.2: Posterior summaries of R_t using the stochastic and deterministic EI-ww models for the Steep, Shallow and Fixed R_t curves. The first column is the simulated data models were fit to. The second column shows the posterior summaries for the deterministic EI-ww model, the third column shows the posterior summaries of the stochastic EI-ww model. First row is the Steep true R_t curve, second row is the Shallow true R_t curve, third row is the Fixed true R_t curve. True R_t values are shown in orange, black lines are posterior medians, blue shaded areas are credible intervals.

For the Steep true R_t curve, both models perform largely the same, although the stochastic

EI-ww model is a little flatter than the deterministic model. For the Shallow and Fixed curves, the deterministic EI-ww model is worse at covering the true shape of the R_t curve.

We summarise the performance of our models across all 100 data sets for each scenarios by reporting some frequentist metrics, shown in Figure 5.3. For each simulation the envelope is a measure of coverage, and is the proportion of time points for which an 80% credible interval from the posterior distribution captured the true value of interest. Mean credible interval width (MCIW) is the mean of 80% credible interval widths across time points within a simulation. Absolute deviation is a measure of accuracy, and is the mean of the absolute difference between the posterior median and the true value at each time point. Finally, mean absolute sequential variation (MASV) measures the variation in the effective reproduction number across time by computing the mean of the absolute difference between the posterior median at t and the posterior median at $t-1$. Each simulation has its own true MASV, which is the difference between the true R_t at t and $t-1$, we summarise the true MASV with its own box-plot, ideally the box-plot of the model's MASV would exactly match the box-plot of the true MASV.

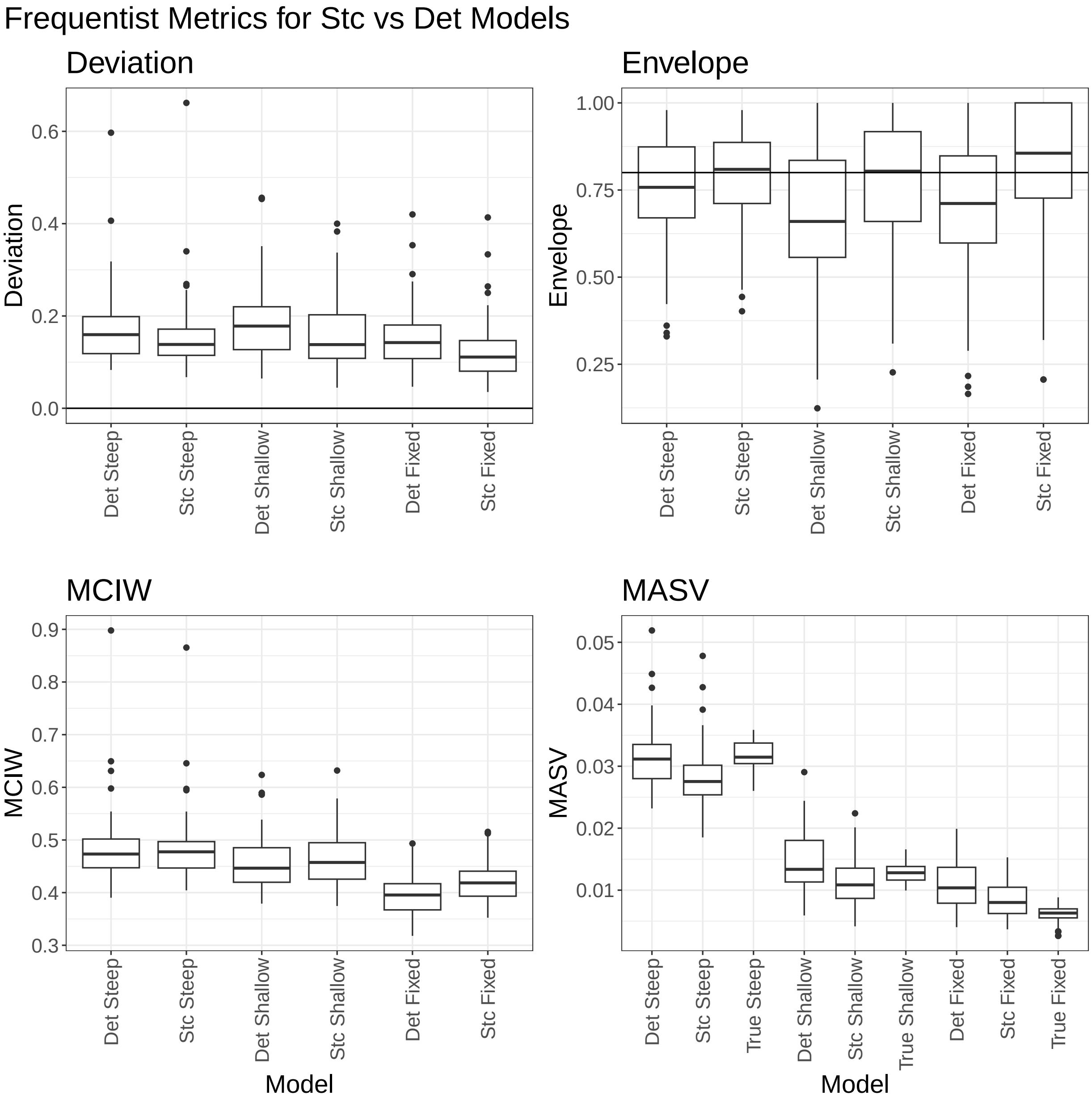


Figure 5.3: Frequentist metrics for the deterministic and stochastic models in the Steep, Shallow and Fixed curve scenarios. For the top row and bottom left plots, the x-axis describes the model and scenario, for the bottom right plot (MASV), the x-axis includes extra values for the true MASV for each simulation scenario. Absolute deviation is the mean of the absolute value of the difference between the median R_t at each time point and the true value. Envelope is a measure of coverage, taking the average coverage of 80% intervals over the time series. MCIW is the average mean credible interval width. Mean absolute standard deviation (MASV) is the difference between the current median point estimate for R_t and the previous point estimate for R_t, we compare it against the true MASV for each simulation.

In all three scenarios, the stochastic model is more accurate and more uncertain. For the Steep scenario, the differences in coverage are marginal, for the Shallow scenario the stochastic model is well calibrated, while the deterministic is not. For the Fixed scenario, the stochastic model is slightly conservative, the deterministic model anti-conservative. Across all scenarios, the stochastic model is less prone to sudden changes than the deterministic model, although in the case of the Steep and Shallow curves, the deterministic model is closer to the true MASV. The stochastic model performs reasonably well across all scenarios, while the deterministic model is sometimes clearly worse.

We wanted to explore how stochasticity in the model affected the differences in performance between the stochastic and deterministic EI-ww models. To this end, we kept the shape of the R_t curve the same but changed the stochasticity of the epidemic by varying the total number of individuals in the population, as well as the number of individuals starting in the E and I compartments. These differences are summarised in Figure 5.4.

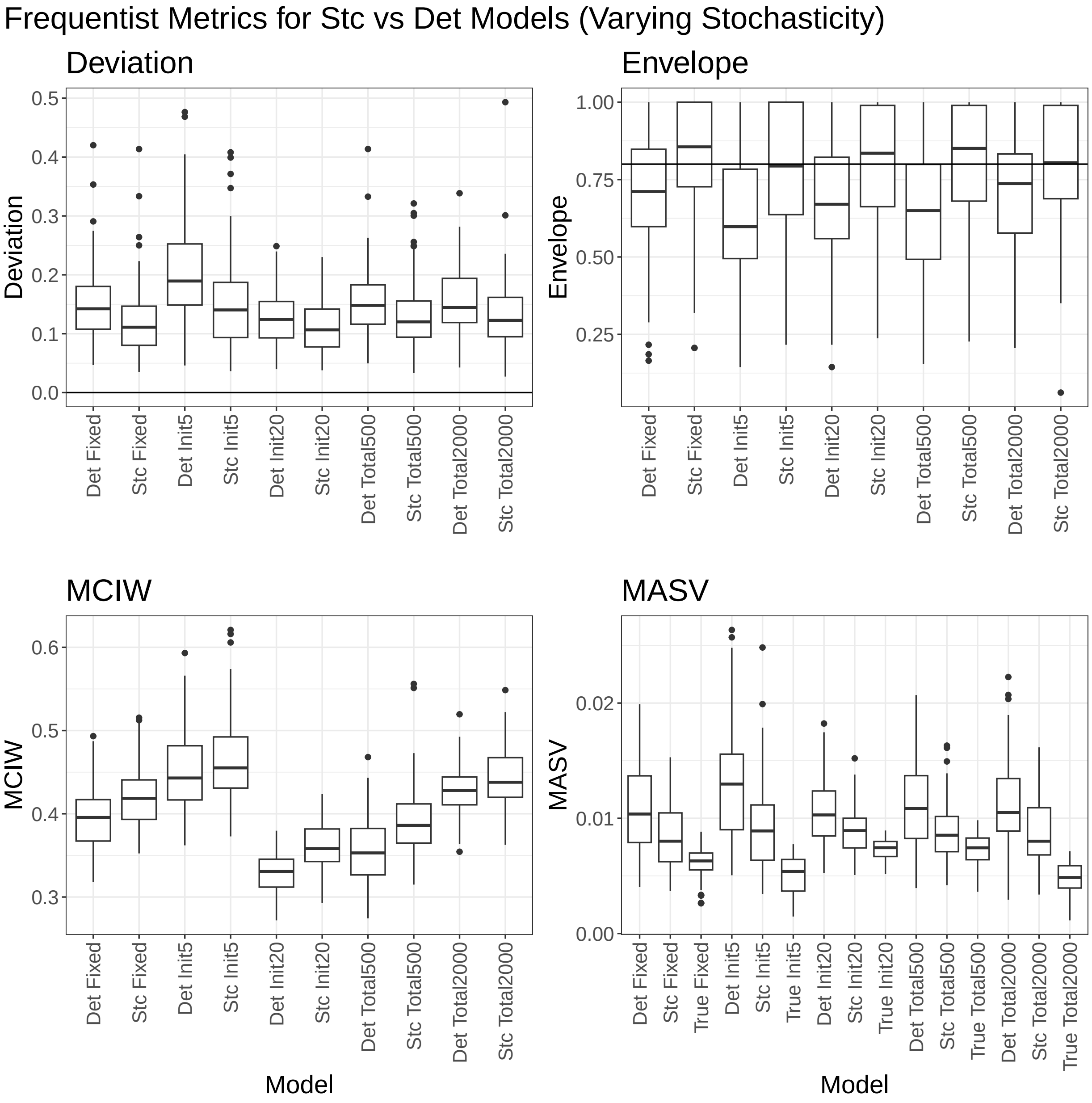


Figure 5.4: Frequentist metrics for the deterministic and stochastic models with varying initial and total populations for the Fixed scenario. For the top row and bottom left plots, the x-axis describes the model and scenario, for the bottom right plot (MASV), the x-axis includes extra values for the true MASV for each simulation scenario. All scenarios are variations on the Fixed scenario with $R_0 = 1.4$. Init5 and Init20 refer to the initial number of individuals in the E and I compartments (5 and 20 respectively), Total500 and Total2000 refer to the total population size (500 and 2000 respectively). The original Fixed scenario with 10 individuals in the E and I compartment and total population size 1000 is included for reference. See Figure 5.3 for metric definitions.

The stochastic EI-ww model is conservative in all scenarios, and the general behaviors we saw in previous scenarios (more accuracy, more uncertainty) remain consistent. The deterministic EI-ww model has noteably worse coverage than the stochastic EI-ww model when the population is only 500 individuals, or when the initial number of individuals is 5 instead of ten. In general, the stochastic model is more conservative and more accurate than the deterministic model.

5.4 The Effective Reproduction Number in UC Irvine Residential Communities

Surveillance data from the SARS-CoV-2 pandemic at the University of California, Irvine were available between January 2022 and June 2022. About 860 wastewater samples were collected from 13 different student housing communities on the University of California Irvine campus from January 2022 to June 2022. These samples were analyzed for SARS-CoV-2 N2 and E genes and water quality parameters such as TSS, COD and ammonia. Usually three replicates were collected per day, if any three was marked as failed, all three were discarded. In addition, counts of new cases were also available. While wastewater data were available at a sub-community level (e.g. individual buildings or spatial regions of a community), case data were only available at the community level.

We chose to analyze three sub-communities with populations around 1000, the total size of the G community is around 3000, the total size of the E community is around 1500. The G community and the E community are spatially not close, both house undergraduate students. While data from December and January were available, we chose not to analyze these data as the campus was on winter break, and then delayed in person classes for the first few weeks of the winter quarter, and so the campus residential population was not stable. We also expect that many reported cases were from individuals returning from travel and testing positive, thus not representing local transmission events. We chose to not analyze data from June 2022 as there were no case data available for June. UCI had a policy of randomly testing individuals in the population which was discontinued in mid-March of 2022. This change in testing policy is accounted for in our case model by modeling the case detection rate as dependent on an indicator variable that equals 0 before the change in policy, and 1 after the change in policy.

We fit both the stochastic and deterministic EI-ww wastewater models, as well as the Epidemia case model to these data sets. The priors for the EI-ww model were the same as in simulations, except for the initial conditions (Appendix Sections C.5.2 and C.7) and initial R_t that was centered around the posterior median of R_t for Orange County, CA (where UC Irvine is located) from a previous analysis using case and test data (Chapter 3). See Appendix Section C.4 for the priors for Epidemia. The data are displayed in Figure 5.5, and the posterior trajectories of R_t are summarised in Figure 5.6.

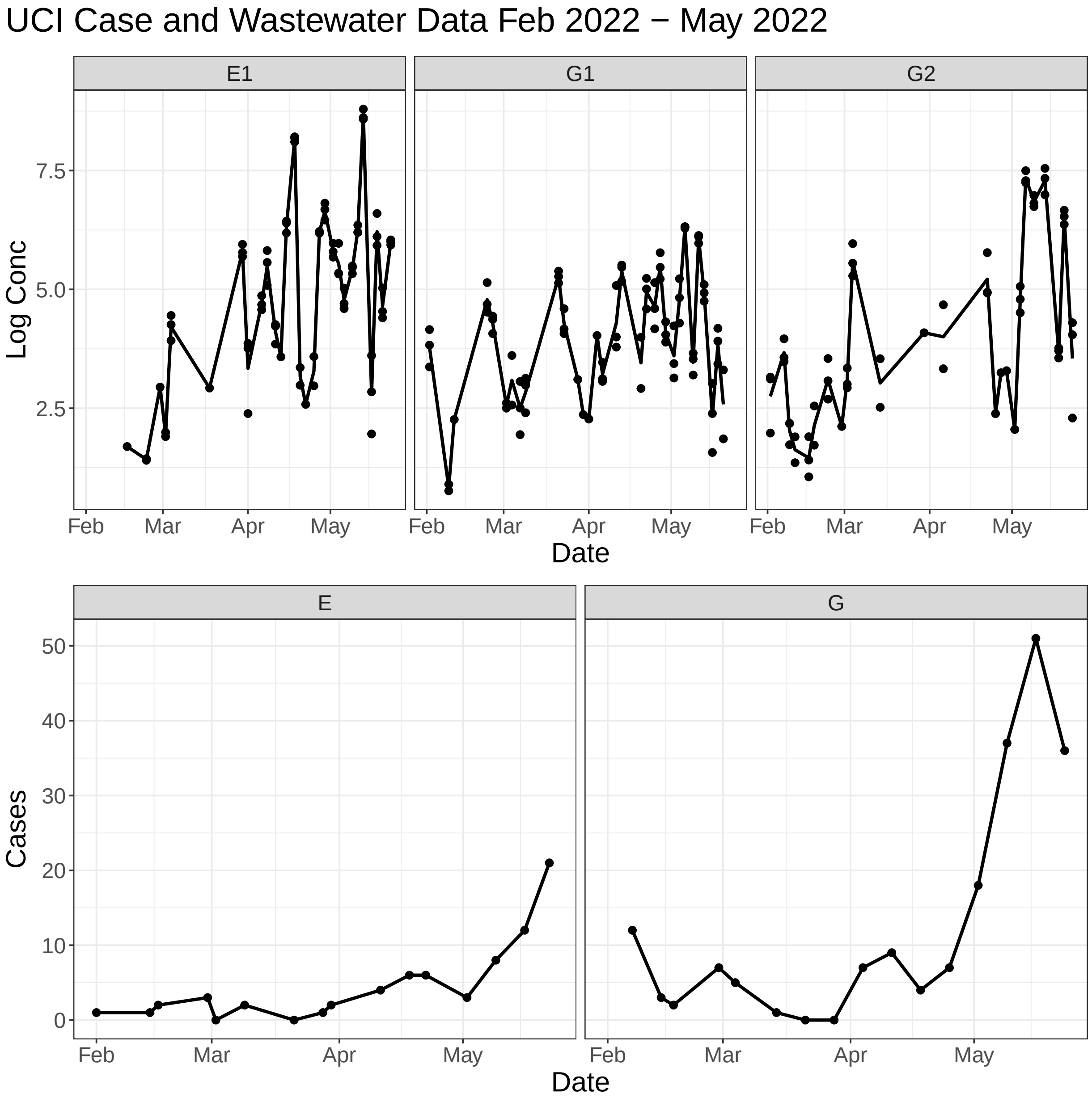


Figure 5.5: Log concentrations of SARS-CoV-2 RNA and weekly reported new COVID-19 cases at UC Irvine for February 2022 through May 2022. For the log concentrations, the dots are individual replicates, and the lines are the mean. The cases are reported at the community level, while the concentrations are reported at the sub-community level. G1 and G2 are sub-communities within the larger G community, while E1 is a sub-community of the E community.

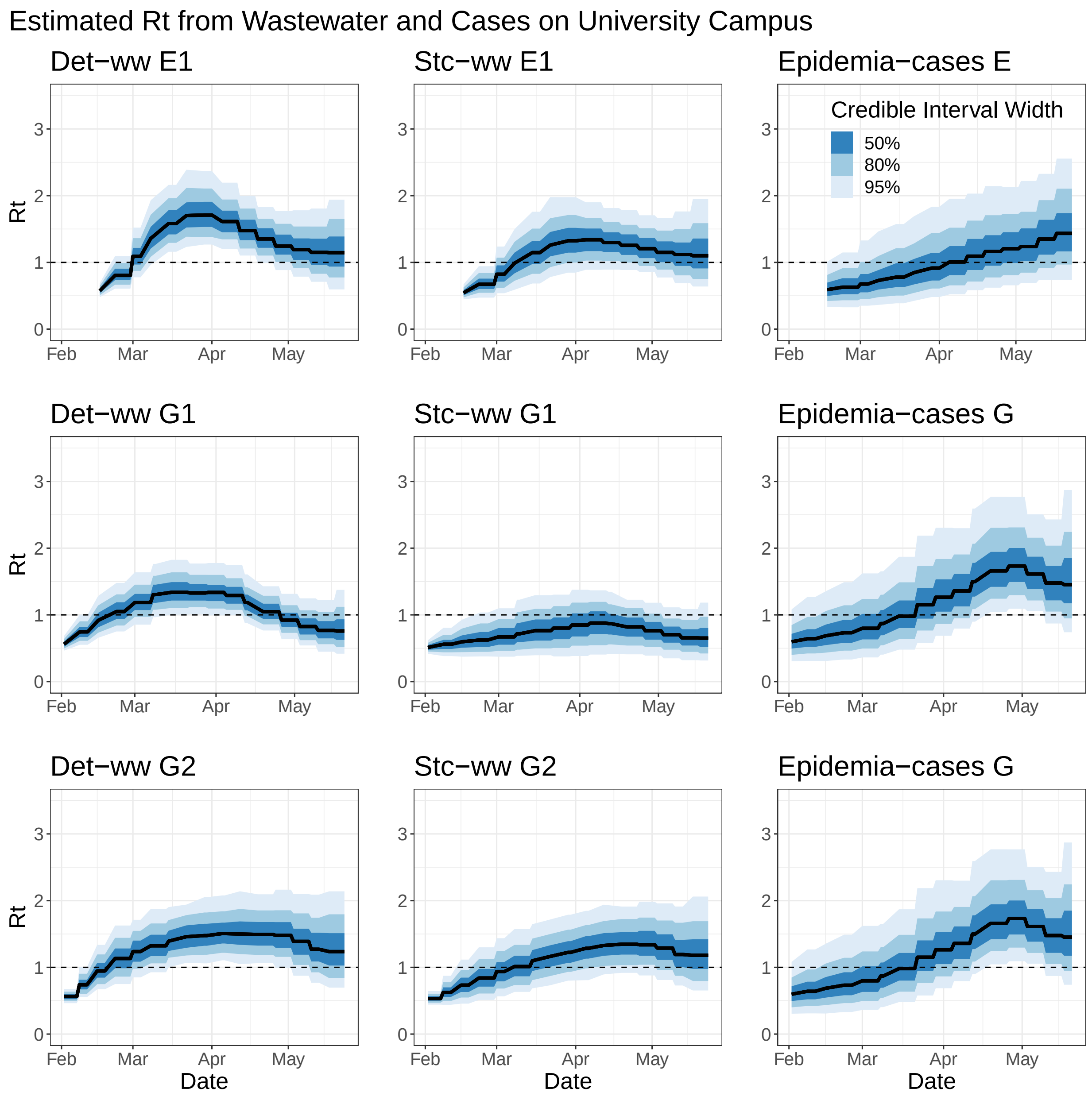


Figure 5.6: Posterior summaries of the effective reproduction number in college campus communities estimated from wastewater data only (stochastic and deterministic models) or case data only (Epidemia-cases). Black lines are posterior medians, blue shaded regions are credible intervals. G1 and G2 are sub-populations of the G campus community, E1 is a sub-population of the E campus community. The bottom two panels in the third column are identical.

The deterministic models show more dramatic changes in R_t than the stochastic models. There is also clear variability in the estimated R_t between different communities and sub-communities, for example, we might reasonably conclude that G2 experienced an outbreak between March and May, while it would be hard to conclude the same about G1. These differences are averaged over in our case-based estimate of the whole community. The wastewater models seem to detect an increase in R_t before the case-based models do in G2 and E1. The posterior medians of the wastewater models are above 1 weeks or months before the posterior medians of the case-based models. The stochastic wastewater-based posterior estimates are still uncertain enough that the 95% credible intervals are never above 1. Figure C.5 shows the results of an alternative analysis using a different prior on the initial value of R_t. While some of the specific conclusions change with the different prior, the overall behavior of our three models is largely consistent with the original analysis.

5.5 Discussion

In this chapter we developed a model with a stochastic latent epidemic process and compared it against the deterministic version in small population settings. Overall, we found the performance of the two models varied based on the shape of the effective reproduction number curve they were trying to estimate, as well as the size of the population experiencing the epidemic. The stochastic EI-ww model was generally more accurate and more uncertain than the deterministic EI-ww model, leading to overall better frequentist calibration. We also used our models to estimate the effective reproduction number of SARS-CoV-2 at UC Irvine, showing qualitatively different estimates between the stochastic and deterministic versions, as well as differences between models using wastewater versus case data.

The stochastic EI-ww model performed generally better than the deterministic EI-ww model, but, depending on the simulation setting, could be conservative. The wide range of envelope values also speaks to highly variable (and in some cases quite poor) performance on individual realizations of the different simulations. We speculate this wide variability in performance arises from the inherently noisy data, and the fact that our model is misspecified, i.e. it does not take into account the time-varying dynamics of pathogen genome shedding [Hoffmann and Alsing, 2023]. We expect this issue of misspecification to be more acute in small populations, where a change in state for a single individual can lead to large variation in overall shedding. In terms of the noisiness of the data, obtaining narrower posterior credible intervals may require using multiple data sources, models that make use of spatial correlation to combine information, or models that are less mis-specified than our current approach (or a combination of all three).

In our real data application, while estimates were often uncertain from wastewater data, they were still qualitatively different than the estimates obtained from case data. The 95% credible intervals from the stochastic EI-ww model always went below one throughout the observation period, on the other hand, in the case of E1 and G2, the curve estimated from wastewater data shows with reasonably high probability R_t was above 1 weeks before the case model. These results suggest using wastewater data in small populations has the promise to provide actionable insights not obtainable from using case data alone.

The deterministic EI-ww model provided similar results to the stochastic EI-ww model, although it inferred stepper increases in R_t and occasionally had credible intervals above 1. Given our experiments, we view the stochastic EI-ww model results as more conservative and more accurate. Depending on the context, public health decision makers may prefer the more conservative model that is less prone to overly dramatic inference.

In addition, our study demonstrates that epidemic dynamics vary based on community and sub-communities. Studies at different levels of population aggregation reveal different dynamics, which in turn require different actions from public health officials. Developing methodologies that continue to improve performance at high spatial resolution seems like a fruitful area of future research.

In our model, we required that the compartment counts be non-zero. This limits our model to situations where we plausibly believe the pathogen is circulating among a non-zero number of individuals. Developing an alternative method that does not require this restriction is an important adaptation of our approach to pursue.

Chapter 6

Discussion and Future Directions

This dissertation describes new methodologies for estimating infectious disease transmission dynamics using novel epidemic models and novel data sources.

In Chapter 3, we developed a method for estimating the effective reproduction number from cases and tests, based on a branching process approximation of the epidemic. The model explicitly allows for the reporting rate to be a function of the total number of tests administered, allowing it to change over time with testing capacity. We showed how this model outperformed state-of-the art methods in simulations, and could be applied to estimate the effective reproduction number of SARS-CoV-2 across the fifteen most populous counties in California. To our knowledge, our approach is one of the first to incorporate total diagnostic tests administered into inference of R_t as a data source. One place where our approach still falls short is that our approach assumes the reporting rate only changes because of total testing volume, where in reality it may change for other reasons such as policies regarding testing access, as well as the population's willingness to test. Modeling such changes over time using case and test data alone is difficult. This challenge directly motivated the method we developed in Chapter 4.

In Chapter 4, we used a modified compartmental model to estimate the effective reproduction number from pathogen genomes samples from wastewater. Our approach modeled wastewater data as realizations of the number of infectious and recently recovered individuals in the population, and ignored the time-varying nature of pathogen shedding. We showed in simulation studies that the model performed well despite this model misspecification, and applied it to estimate the effective reproduction number of SARS-CoV-2 in Los Angeles, California. We compared the estimated effective reproduction number with estimates from models using case data, and showed that they agreed in many crucial places, but also disagreed at times, highlighting the potential uses of wastewater data to provide novel insights into epidemic outbreaks. We also showed how compartmental models could be adapted to exclude the susceptible compartment, and connected with either case or wastewater data. This framework has the potential to be used with many kinds of data. However, our method relied on a deterministic compartmental model, which is not always suitable in small population settings.

Chapter 5 adapted the method developed in Chapter 4 to estimate the effective reproduction number to the small population setting. To do this, we developed a stochastic compartmental model with Gaussian transition densities and closed form conditional moments derived from the Markov Jump Process representation of the compartmental model. We showed in simulations how this stochastic model performs better than its deterministic counterpart, though it is often conservative. We then applied both deterministic and stochastic models to SARS-CoV-2 wastewater surveillance data collected at several UC Irvine dormitories, demonstrating qualitative differences between the two models, as well as case based estimates of the effective reproduction number for the same setting.

There are a few obvious opportunities for future projects building on the work described in this thesis. Perhaps most naturally, we have fit models to case and wastewater data separately, but have not yet fit a model to both data sources jointly. It is likely a joint model would allow us to model the reporting rate as time-varying, addressing a major limitation of our case model from Chapter 3. Also, since case data is less noisy than wastewater data, we expect we would get tighter credible intervals than those produced by the models in Chapters 4 and 5. As a first step, the model would be a relatively straightforward combination of the EIR and EIRR models described in Chapter 4, however, to account for problems in changing case detection rates, the model would use a time-varying case detection parameter ψ_t modeled using a random walk prior similar to the random walk prior used for the effective reproduction number.

$$\begin{aligned} P(\mathbf{R}, \bar{\boldsymbol{\Theta}}, \rho, \lambda, \tau, df, \sigma, \boldsymbol{\psi}, \sigma_\psi, \phi, \mid \mathbf{X}, \mathbf{O}) \propto & \underbrace{P(\mathbf{X} \mid \mathbf{M}(t, \bar{\boldsymbol{\Theta}}, \mathbf{R}), \rho, \lambda, \tau, df)}_{\text{Concentration Model}} \underbrace{P(\mathbf{O} \mid \mathbf{M}(t, \bar{\boldsymbol{\Theta}}, \mathbf{R}), \boldsymbol{\psi}, \phi)}_{\text{Case Model}} \\ & \underbrace{P(\mathbf{R} \mid \sigma)}_{\text{RW Prior}} \underbrace{P(\boldsymbol{\psi} \mid \sigma_\psi)}_{\text{RW Prior}} P(\bar{\boldsymbol{\Theta}}, \rho, \lambda, \tau, df, \sigma, \sigma_\psi, \phi) \end{aligned}$$

where $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_T\}$. The case observation model from Chapter 3 could also be used in this joint model. The idea here is that by using the wastewater data, the the time-varying case detection rate will be identifiable, allowing the case data to usefully contribute to effective reproduction number inference. Data integration is a non-trivial task with many implementation challenges, but is certainly worth pursuing [Bayer et al., 2024].

We also repeatedly inferred the effective reproduction number in multiple spatial locations, without modeling those spatial locations jointly. In reality of course, we expect that neighboring counties, or neighboring dormitories, are spatially correlated. Teh et al. [2022] show one way to explicitly model spatial correlation, by using a Gaussian process prior where covariance depends on both spatial location and time, as well as explicitly allowing for infectious individuals to infect individuals in different spatial locations based on commuting data. In the compartmental modeling context, one can explicitly model infections across spatial locations through the rate of new infections. Suppose we have two spatial locations (areas 1 and 2), each with their own set of SEIR compartments and populations N_1 and N_2.

The rate of new infections into the E compartment can be written as

$$\text{‘}\left(\frac{\beta_1}{N_1}I_1+\frac{\beta_2}{N_2}I_2\right)S_1,$$

where the term $\frac{\beta_2}{N_2}I_2S_1$ represents the portion of the rate of new infections in area 1 due to the number of infectious individuals in area 2. Fintzi et al. [2022] implemented such a model when analyzing the 2014 Ebola outbreak in three countries.

We have also used the random walk prior to model R_t through all three chapters. This choice is made for expediency, but an important limitation of the random walk prior is that it requires the standard deviation of the random walk to be constant across time. Allowing this to vary, either through a more flexible stochastic process prior, such as a horseshoe random walk prior, is a promising way to adapt the models described in this dissertation [Faulkner and Minin, 2018, Abbott et al., 2020, Bouman et al., 2023].

In Chapters 4 and 5 we used simple models to describe the relationship between wastewater data and the compartment counts. We know these models are incorrect, and they could be adjusted to allow for more complexity (for example, Nourbakhsh et al. [2022] uses multiple infectious compartments and has an explicit model of sewer dynamics), or by explicitly defining a stochastic model that describes time-varying shedding, and deriving a relationship between compartment counts and pathogen shedding from that model.

Finally, in Chapter 5, we only created a stochastic model for wastewater. The same could be done for cases, although re-parametrization of the MJP is required in order to ensure that for the diffusion approximation the total number of newly infected individuals never decreases. To modify our model to be fit to case data, we re-parametrize in the manner of Fintzi et al. [2022], so that the model now counts the cumulative transitions into compartments as opposed to the counts in the compartments themselves. The original MJP $\mathbf{H}(t) = (E(t), I(t))$ counts the number of infected but not infectious and currently infectious individuals. Let

$\mathbf{J}(t) = (L(t), M(t), N(t))^T$ be the cumulative number of transitions into the E comparment, into the I compartment, and out of the I compartment respectively. Then

$$\mathbf{H}(t) = \mathbf{H}(0) + \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \mathbf{J}(t). \tag{6.1}$$

We then use the same technique described in Chapter 5 to solve for the first and second moments of $\mathbf{J}(t)$, conditioned on $\mathbf{H}(0) = (e_0, i_0)$. The resulting differential equations are now inhomogeneous, but still linear, and their solutions are still available in closed form (albeit a more complex one).

The rest of the model is the same as the case model used in Chapter 4. The difficulty with implementing this model is simply that the closed form solution of the moments is more complex than the previous model. It may be better to use a numerical solver instead.

Bibliography

Sam Abbott, Joel Hellewell, Robin N Thompson, Katharine Sherratt, Hamish P Gibbs, Nikos I Bosse, James D Munday, Sophie Meakin, Emma L Doughty, June Young Chun, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 2; peer review: 1 approved with reservations]. *Wellcome Open Res 2020, 5:112*, 5(112):112, 2020. URL `https://doi.org/10.12688/wellcomeopenres.16006.2`.

Sam Abbott, Katharine Sherratt, Moritz Gerstung, and Sebastian Funk. Estimation of the test to test distribution as a proxy for generation interval distribution for the Omicron variant in England. *medRxiv*, page 2022.01.08.22268920, 2022. doi: 10.1101/2022.01.08.22268920.

Patrick T Acer, Lauren M Kelly, Andrew A Lover, and Caitlyn S Butler. Quantifying the relationship between SARS-CoV-2 wastewater concentrations and building-level COVID-19 prevalence at an isolation residence: a passive sampling approach. *International Journal of Environmental Research and Public Health*, 19(18):11245, 2022.

Linda JS Allen. *An Introduction to Stochastic Processes with Applications to Biology*. CRC press, 2010.

Hakan Andersson and Tom Britton. *Stochastic Epidemic Models and their Statistical Analysis*, volume 151. Springer Science & Business Media, 2012.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342, 2010.

Frank Ball and Peter Donnelly. Strong approximations for epidemic models. *Stochastic Processes and their Applications*, 55(1):1–21, 1995.

Frank Ball and David Siri. Stochastic SIR models in Structured Populations. In Tom Britton and Etienne Pardoux, editors, *Stochastic Epidemic Models with Inference*, Lecture Notes in Mathematics (LNM), volume 2255, pages 121–157. Springer, 2019. ISBN 978-3-030-30900-8. doi: 10.1007/978-3-030-30900-8.

Andrew D Barbour. On a functional central limit theorem for Markov population processes. *Advances in Applied Probability*, 6(1):21–39, 1974.

Damon Bayer, Isaac H Goldstein, Jonathan Fintzi, Keith Lumbard, Emily Ricotta, Sarah Warner, Lindsay M Busch, Jeffrey R Strich, Daniel S Chertow, Daniel M Parker, et al. Semi-parametric modeling of SARS-CoV-2 transmission using tests, cases, deaths, and seroprevalence data. *Annals of Applied Statistics*, 2024.

Amy E Benefield, Laura A Skrip, Andrea Clement, Rachel A Althouse, Stewart Chang, and Benjamin M Althouse. SARS-CoV-2 viral load peaks prior to symptom onset: a systematic review and individual-pooled analysis of coronavirus viral load from 66 studies. *medrxiv*, 2020.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Samir Bhatt, Neil Ferguson, Seth Flaxman, Axel Gandy, Swapnil Mishra, and James A Scott. Semi-mechanistic Bayesian modeling of COVID-19 with renewal processes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page in press, 02 2023. ISSN 0964-1998. doi: 10.1093/jrsssa/qnad030.

Joris Bierkens and Gareth Roberts. A piecewise deterministic scaling limit of lifted Metropolis-Hastings in the Curie-Weiss model. *The Annals of Applied Probability*, 27 (2):846 – 882, 2017.

Julie C Blackwood and Lauren M Childs. An introduction to compartmental modeling for the budding infectious disease modeler. *Letters in Biomathematics*, 5(1):195 – 221, 2018.

Judith Bouman, Anthony Hauser, Simon L Grimm, Martin Wohlfender, Samir Bhatt, Elizaveta Semenova, Andrew Gelman, Christian L Althaus, and Julien Riou. Bayesian workflow for time-varying transmission in stratified compartmental infectious disease transmission models. *medRxiv*, pages 2023–10, 2023.

Tom Britton and Etienne Pardoux. Stochastic Epidemics in a Homogenous Community. In Tom Britton and Etienne Pardoux, editors, *Stochastic Epidemic Models with Inference*, Lecture Notes in Mathematics (LNM), volume 2255, pages 3–119. Springer, 2019. ISBN 978-3-030-30900-8. doi: 10.1007/978-3-030-30900-8.

Elizabeth Buckingham-Jeffery, Valerie Isham, and Thomas House. Gaussian process approximations for fast inference from infectious disease data. *Mathematical Biosciences*, 301: 111–120, 2018.

Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

California Open Data Portal. California Open Data Portal. `https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state`, 2023. [Online; accessed 10-June-2023].

Marcos A Capistrán, Antonio Capella, and J Andrés Christen. Filtering and improved uncertainty quantification in the dynamic estimation of effective reproduction numbers. *Epidemics*, 40:100624, 2022.

Simon Cauchemez and Neil M Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society Interface*, 5(25):885–897, 2008.

David Champredon and Jonathan Dushoff. Intrinsic and realized generation intervals in infectious-disease transmission. *Proceedings of the Royal Society B: Biological Sciences*, 282(1821):2015–2026, 2015. doi: 10.1098/rspb.2015.2026.

David Champredon, Jonathan Dushoff, and David J. D. Earn. Equivalence of the Erlang-distributed SEIR epidemic model and the renewal equation. *SIAM Journal on Applied Mathematics*, 78(6):3258–3278, 2018. doi: 10.1137/18M1186411.

Alice Corbella, Trevelyan J McKinley, Paul J Birrell, Anne M Presanis, Simon EF Spencer, Gareth O Roberts, and Daniela De Angelis. The lifebelt particle filter for robust estimation from low-valued count data. *arXiv preprint arXiv:2212.04400*, 2022a.

Alice Corbella, Simon EF Spencer, and Gareth O Roberts. Automatic Zig-Zag sampling in practice. *Statistics and Computing*, 32(6):107, 2022b.

Anne Cori, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, 2013. ISSN 0002-9262. doi: 10.1093/aje/kwt133. URL `https://academic.oup.com/aje/article/178/9/1505/89262`.

Kenny Crump and Charles J Mode. A general age-dependent branching process. I. *Journal of Mathematical Analysis and Applications*, 24:494–508, 1968.

Kenny Crump and Charles J Mode. A general age-dependent branching process. II. *Journal of Mathematical Analysis and Applications*, 25(1):8–17, 1969.

Odo Diekmann, JAP Heesterbeek, and Michael G Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885, 2010.

Peter Diggle, Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger, et al. *Analysis of Longitudinal Data*. Oxford university press, 2002.

Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain. fields: Tools for spatial data, 2021. URL `https://github.com/dnychka/fieldsRPackage`. R package version 14.1.

Claire Duvallet, Fuqing Wu, Kyle A McElroy, Maxim Imakaev, Noriko Endo, Amy Xiao, Jianbo Zhang, Róisín Floyd-O'Sullivan, Morgan M Powell, Samuel Mendola, et al. Nationwide trends in COVID-19 cases and SARS-CoV-2 RNA wastewater concentrations in the United States. *ACS Es&t Water*, 2(11):1899–1909, 2022.

James R Faulkner and Vladimir N Minin. Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis*, 13(1):225, 2018.

William Feller. *An Introduction to Probability Theory and Its Applications: 2nd edition*, volume 1. John Wiley and Sons, 1957.

Luca Ferretti, Alice Ledda, Chris Wymant, Lele Zhao, Virginia Ledda, Lucie Abeler-Dörner, Michelle Kendall, Anel Nurtay, Hao-Yuan Cheng, Ta-Chou Ng, Hsien-Ho Lin, Rob Hinch, Joanna Masel, A. Marm Kilpatrick, and Christophe Fraser. The timing of COVID-19 transmission. *medRxiv*, 2020. doi: 10.1101/2020.09.04.20188516.

Bärbel F Finkenstädt and Bryan T Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 49(2):187–205, 2000.

Jonathan Fintzi, Xiang Cui, Jon Wakefield, and Vladimir N Minin. Efficient data augmentation for fitting stochastic epidemic models to prevalence data. *Journal of Computational and Graphical Statistics*, 26(4):918–929, 2017.

Jonathan Fintzi, Jon Wakefield, and Vladimir N Minin. A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. *Biometrics*, 78(4): 1530–1541, 2022.

Jonathan Refael Fintzi. *Bayesian Modeling of Partially Observed Epidemic Count Data*. PhD thesis, University of Washington, 2018.

Seth Flaxman, Swapnil Mishra, and Axel et. al. Gandy. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2405-7. URL `https://doi.org/10.1038/s41586-020-2405-7`.

Christopher Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLOS ONE*, 2(8):1–12, 2007. doi: 10.1371/journal.pone.0000758. URL `https://doi.org/10.1371/journal.pone.0000758`.

Jerome Friedman, Robert Tibshirani, and Trevor Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Christiane Fuchs. *Inference for Diffusion Processes: with Applications in Life Sciences*. Springer Science & Business Media, 2013.

Tapiwa Ganyani, Cecile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*, 25(17):2000257, 2020.

Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: A language for flexible probabilistic inference. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1682–1690. PMLR, 09–11 Apr 2018.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2014.

Gavin J Gibson and Eric Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology: A Journal of the IMA*, 15(1):19–40, 1998.

Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

Isaac H Goldstein. Rt estimates for California counties. `https://github.com/igoldsteinh/CA_rt_estimates`, 2024. [Online; accessed 05-May-2024].

Isaac H Goldstein, Jon Wakefield, and Volodymyr M Minin. Incorporating testing volume into estimation of effective reproduction number dynamics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(2):436–453, 2024.

Andrew Golightly, Laura E Wadkin, Sam A Whitaker, Andrew W Baggaley, Nick G Parker, and Theodore Kypraios. Accelerating Bayesian inference for stochastic epidemic models using incidence data. *Statistics and Computing*, 33(6):134, 2023.

Katelyn M. Gostic, Lauren McGough, Edward B. Baskerville, and Sam et. al. Abbott. Practical considerations for measuring the effective reproductive number, Rt. *PLOS Computational Biology*, 16(12):1–21, 2020. doi: 10.1371/journal.pcbi.1008409.

Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29, 2020. doi: 10.18637/jss.v092.i10.

Mi Seon Han, Moon-Woo Seong, Namhee Kim, Sue Shin, Sung Im Cho, Hyunwoong Park, Taek Soo Kim, Sung Sup Park, and Eun Hwa Choi. Viral RNA load in mildly symptomatic and asymptomatic children with COVID-19, Seoul, South Korea. *Emerging infectious diseases*, 26(10):2497, 2020.

William S Hart, Elizabeth Miller, Nick J Andrews, Pauline Waight, Philip K Maini, Sebastian Funk, and Robin N Thompson. Generation time of the Alpha and Delta SARS-CoV-2 variants: an epidemiological analysis. *The Lancet Infectious Diseases*, 22(5):603–610, 2022.

Luke S Hillary, Shelagh K Malham, James E McDonald, and David L Jones. Wastewater and public health: the potential of wastewater surveillance for monitoring COVID-19. *Current Opinion in Environmental Science & Health*, 17:14–20, 2020.

Lam Si Tung Ho, Forrest W Crawford, and Marc A Suchard. Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. *The Annals of Applied Statistics*, 12(3):1993–2021, 2018.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47): 1593–1623, 2014.

Till Hoffmann and Justin Alsing. Faecal shedding models for SARS-CoV-2 RNA among hospitalised patients and implications for wastewater-based epidemiology. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(2):330–345, 2023.

Jana S Huisman, Jérémie Scire, Daniel C Angst, Jinzhou Li, Richard A Neher, Marloes H Maathuis, Sebastian Bonhoeffer, and Tanja Stadler. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *Elife*, 11:e71345, 2022a.

Jana S Huisman, Jérémie Scire, Lea Caduff, Xavier Fernandez-Cassi, Pravin Ganesanandamoorthy, Anina Kull, Andreas Scheidegger, Elyse Stachler, Alexandria B Boehm, Bridgette Hughes, et al. Wastewater-based estimation of the effective reproductive number of SARS-CoV-2. *Environmental Health Perspectives*, 130(5):057011, 2022b.

Wolfram Research, Inc. Mathematica, Version 13.1. URL `https://www.wolfram.com/mathematica`. Champaign, IL, 2022.

Valerie Isham. Assessing the variability of stochastic epidemics. *Mathematical Biosciences*, 107(2):209–224, 1991.

Peter Jagers. *Branching Processes with Biological Applications.* Wiley, 1975.

Nicholas P Jewell and Joseph A Lewnard. On the use of the reproduction number for SARS-CoV-2: Estimation, misinterpretations and relationships with other ecological measures. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185:S16–S27, 2022.

Samuel Karlin. *A First Course in Stochastic Processes.* Academic Press, 1966.

James W Keck, Reuben Adatorwovor, Matthew Liversedge, Blazan Mijotavich, Cullen Olsson, William D Strike, Atena Amirsoleimani, Ann Noble, Soroosh Torabi, Alexus Rockward, et al. Wastewater surveillance for identifying SARS-CoV-2 infections in long-term care facilities, Kentucky, USA, 2021–2022. *Emerging Infectious Diseases*, 30(3):530, 2024.

Matt J Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals.* Princeton University Press, 2008.

Ben Killingley, Alex J Mann, Mariya Kalinova, Alison Boyers, Niluka Goonawardane, Jie Zhou, Kate Lindsell, Samanjit S Hare, Jonathan Brown, Rebecca Frise, et al. Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge in young adults. *Nature Medicine*, 28(5):1031–1041, 2022.

Marek Kimmel. The point-process approach to age-and time-dependent branching processes. *Advances in Applied Probability*, 15(1):1–20, 1983.

Aaron A King, Edward L Ionides, and Jesse Wheeler. Simulation-based Inference for Epidemiological Dynamics . `https://kingaa.github.io/sbied/`. [Online; accessed 14-Apr-2024].

Aaron A. King, Dao Nguyen, and Edward L. Ionides. Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*, 69 (12):1–43, 2016. doi: 10.18637/jss.v069.i12.

Thomas G Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

Thomas G Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356, 1971.

Theodore Kypraios and Philip D O'Neill. MCMCII for infectious diseases. `https://www.maths.nottingham.ac.uk/plp/pmztk/files/MCMC2-Seattle/`, 2021. [Online; accessed 14-Apr-2024].

Xintong Li, Howard H Chang, Qu Cheng, Philip A Collender, Ting Li, Jinge He, Lance A Waller, Benjamin A Lopman, and Justin V Remais. A spatial hierarchical model for integrating and bias-correcting data from passive and active disease surveillance systems. *Spatial and Spatio-Temporal Epidemiology*, 35:100341, 2020.

Grace Lui, Lowell Ling, Christopher KC Lai, Eugene YK Tso, Kitty SC Fung, Veronica Chan, Tracy Hy Ho, Fion Luk, Zigui Chen, Joyce KC Ng, et al. Viral dynamics of SARS-CoV-2 across a spectrum of disease severity in COVID-19. *Journal of Infection*, 81(2): 318–356, 2020.

Rasha Maal-Bared, Yuanyuan Qiu, Qiaozhi Li, Tiejun Gao, Steve E Hrudey, Sudha Bhavanam, Norma J Ruecker, Erik Ellehoj, Bonita E Lee, and Xiaoli Pang. Does normalization of SARS-CoV-2 concentrations by Pepper Mild Mottle Virus improve correlations and lead time between wastewater surveillance and clinical data in Alberta (Canada): comparing twelve SARS-CoV-2 normalization approaches. *Science of The Total Environment*, 856: 158964, 2023.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Swapnil Mishra, Jamie Scott, Harrison Zhu, Neil M. Ferguson, Samir Bhatt, Seth Flaxman, and Axel Gandy. A COVID-19 model for local authorities of the United Kingdom. *medRxiv*, 2020. doi: 10.1101/2020.11.24.20236661.

Fuminari Miura, Masaaki Kitajima, and Ryosuke Omori. Duration of SARS-CoV-2 viral shedding in faeces as a parameter for wastewater-based epidemiology: Re-analysis of patient data using a shedding dynamics model. *Science of The Total Environment*, 769: 144549, 2021.

Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003. doi: 10.1137/S00361445024180. URL `https://doi.org/10.1137/S00361445024180`.

Raphael Morsomme and Jason Xu. Exact inference for stochastic epidemic models via uniformly ergodic block sampling. *arXiv preprint arXiv:2201.09722*, 2022.

Mario Morvan, Anna Lo Jacomo, Celia Souque, Matthew J Wade, Till Hoffmann, Koen Pouwels, Chris Lilley, Andrew C Singer, Jonathan Porter, Nicholas P Evens, et al. An analysis of 45 large-scale wastewater sites in England to estimate SARS-CoV-2 community prevalence. *Nature Communications*, 13(1):4313, 2022.

Rebecca K. Nash, Pierre Nouvellet, and Anne Cori. Real-time estimation of the epidemic reproduction number: Scoping review of the applications and challenges. *PLOS Digital Health*, 1(6):1–17, 2022.

Peter J Neal. Approximate Bayesian computation methods for epidemic models. In Leonard Held, Niels Hens, Phil O'Neill, and Jacco Wallinga, editors, *Handbook of Infectious Disease Data Analysis*, Handbooks of Modern Statistical Methods, pages 179–195. Chapman and Hall/CRC, 2020.

Radford M Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L Jones, and Xioa-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, pages 179–195. Chapman and Hall/CRC, 2011.

Shokoofeh Nourbakhsh, Aamir Fazil, Michael Li, Chand S Mangat, Shelley W Peterson, Jade Daigle, Stacie Langner, Jayson Shurgold, Patrick D'Aoust, Robert Delatolla, et al. A wastewater-based epidemic model for SARS-CoV-2 with application to three Canadian cities. *Epidemics*, 39:100560, 2022.

Yasutaka Okita, Takayoshi Morita, and Atsushi Kumanogoh. Duration of SARS-CoV-2 RNA positivity from various specimens and clinical characteristics in patients with COVID-19: a systematic review and meta-analysis. *Inflammation and Regeneration*, 42(1):1–19, 2022.

Bernt Oksendal. *Stochastic Differential Equations: an Introduction with Applications*. Springer Science & Business Media, 2013.

Philip D O'Neill and Gareth O Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 162(1): 121–129, 1999.

Mikko S Pakkanen, Xenia Miscouridou, Matthew J Penn, Charles Whittaker, Tresnia Berah, Swapnil Mishra, Thomas A Mellan, and Samir Bhatt. Unifying incidence and prevalence under a time-varying general branching process. *Journal of Mathematical Biology*, 87(2): 35, 2023.

Kris V Parag. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. *PLoS Computational Biology*, 17(9):e1009347, 2021.

Sang Woo Park, Kaiyuan Sun, David Champredon, Michael Li, Benjamin M. Bolker, David J. D. Earn, Joshua S. Weitz, Bryan T. Grenfell, and Jonathan Dushoff. Forward-looking serial intervals correctly link epidemic growth to reproduction numbers. *Proceedings of the National Academy of Sciences*, 118(2):e2011548118, 2021. ISSN 0027-8424. doi: 10.1073/pnas.2011548118.

Matthew J Penn, Daniel J Laydon, Joseph Penn, Charles Whittaker, Christian Morgenstern, Oliver Ratmann, Swapnil Mishra, Mikko S Pakkanen, Christl A Donnelly, and Samir

Bhatt. The uncertainty of infectious disease outbreaks is underestimated. *arXiv preprint arXiv:2210.14221*, 2022.

David Polo, Marcos Quintela-Baluja, Alexander Corbishley, Davey L Jones, Andrew C Singer, David W Graham, and Jesús L Romalde. Making waves: Wastewater-based epidemiology for COVID-19–approaches and challenges for surveillance and prediction. *Water Research*, 186:116404, 2020.

Koen B Pouwels, Thomas House, Emma Pritchard, Julie V Robotham, Paul J Birrell, Andrew Gelman, Karina-Doris Vihta, Nikola Bowers, Ian Boreham, Heledd Thomas, et al. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS coronavirus infection survey. *The Lancet Public Health*, 6(1):e30–e38, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Christopher Rackauckas and Qing Nie. Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in Julia. *Journal of Open Research Software*, 5(1):15, 2017.

Eric Renshaw. *Stochastic Population Processes: Analysis, Approximations, Simulations*. OUP Oxford, 2015.

Sheldon M Ross. *Introduction to Probability Models*. Academic press, 1985.

Kevin Rupp, Rudolf Schill, Jonas Süskind, Peter Georg, Maren Klever, Andreas Lösch, Lars Grasedyck, Tilo Wettig, and Rainer Spang. Differentiated uniformization: A new method for inferring Markov chains on combinatorial state spaces including stochastic epidemic models. *Computational Statistics*, pages 1–21, 2024.

James A. Scott, Axel Gandy, Swapnil Mishra, Samir Bhatt, Seth Flaxman, H. Juliette T. Unwin, and Jonathan Ish-Horowicz. Epidemia: An R package for semi-mechanistic Bayesian modelling of infectious diseases using point processes. *arXiv*, 2021. doi: 10.48550/ARXIV.2110.12461.

Ron Sender, Yinon M. Bar-On, Sang Woo Park, Elad Noor, Jonathan Dushoff, and Ron Milo. The unmitigated profile of COVID-19 infectiousness. *medRxiv*, 2021. doi: 10.1101/2021.11.17.21266051. URL `https://www.medrxiv.org/content/early/2021/11/25/2021.11.17.21266051`.

Katharine Sherratt, Sam Abbott, Sophie R Meakin, Joel Hellewell, James D Munday, Nikos Bosse, CMMID Covid-19 working group, Mark Jit, and Sebastian Funk. Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England. *Philosophical Transactions of the Royal Society B*, 376(1829):20200283, 2021.

Zhiquan Song, Ryan Reinke, Mike Hoxsey, James Jackson, Eric Krikorian, Nikos Melitas, Diego Rosso, and Sunny Jiang. Detection of SARS-CoV-2 in wastewater: Community variability, temporal dynamics, and genotype diversity. *Acs Es&T Water*, 1(8):1816–1825, 2021.

Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.

Stan Development Team. RStan: the R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

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

Swiss National Covid-19 Science Task Force. Situation report: Reproductive number, 2020. URL `https://ncs-tf.ch/en/situation-report`. [Online; accessed 2020-09-17].

Yee Whye Teh, Bryn Elesedy, Bobby He, Michael Hutchinson, Sheheryar Zaidi, Avishkar Bhoopchand, Ulrich Paquet, Nenad Tomasev, Jonathan Read, and Peter J Diggle. Efficient Bayesian inference of instantaneous reproduction numbers at fine spatial scales, with an application to mapping and nowcasting the COVID-19 epidemic in British local authorities. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185 (Supplement_1):S65–S85, 2022.

Robin N Thompson, Jake E Stockwin, Rolina D van Gaalen, Jonny A Polonsky, Zhian N Kamvar, P Alex Demarsh, Elisabeth Dahlqwist, Siyang Li, Eve Miguel, Thibaut Jombart, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29:100356, 2019.

Juho Timonen, Nikolas Siccha, Ben Bales, Harri Lähdesmäki, and Aki Vehtari. An importance sampling approach for reliable and efficient inference in Bayesian ordinary differential equation models. *arXiv preprint arXiv:2205.09059*, 2022.

Ch Tsitouras. Runge–Kutta pairs of order 5 (4) satisfying only the first column simplifying assumption. *Computers & Mathematics with Applications*, 62(2):770–775, 2011.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

Matthew J Wade, Anna Lo Jacomo, Elena Armenise, Mathew R Brown, Joshua T Bunce, Graeme J Cameron, Zhou Fang, Deidre F Gilpin, David W Graham, Jasmine MS Grimsley, et al. Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: Lessons learned from the United Kingdom national COVID-19 surveillance programmes. *Journal of hazardous materials*, 424:127456, 2022.

Jon Wakefield, Tracy Qi Dong, and Volodymyr N Minin. Spatio-temporal analysis of surveillance data. In Leonard Held, Niels Hens, Phil O'Neill, and Jacco Wallinga, editors, *Handbook of Infectious Disease Data Analysis*, Handbooks of Modern Statistical Methods, pages 455–472. Chapman and Hall/CRC, 2020.

EWJ Wallace, DT Gillespie, KR Sanft, and LR Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET Systems Biology*, 6(4):102–115, 2012.

Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6):509–516, 2004.

Kieran A Walsh, Karen Jordan, Barbara Clyne, Daniela Rohde, Linda Drummond, Paula Byrne, Susan Ahern, Paul G Carty, Kirsty K O'Brien, Eamon O'Murchu, et al. SARS-CoV-2 detection, viral load and infectivity over the course of an infection. *Journal of Infection*, 81(3):357–371, 2020.

WastewaterSCAN. WastewaterSCAN. `https://wastewaterscan.org`, 2023. [Online; accessed 9-May-2023].

Roman Wölfel, Victor M Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A Müller, Daniela Niemeyer, Terry C Jones, Patrick Vollmar, Camilla Rothe, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature*, 581(7809):465–469, 2020.

S.N. Wood. *Generalized Additive Models: An Introduction with R (2nd ed.)*. Chapman and Hall/CRC, 2017.

Hualei Xin, Yu Li, Peng Wu, Zhili Li, Eric HY Lau, Ying Qin, Liping Wang, Benjamin J Cowling, Tim K Tsang, and Zhongjie Li. Estimating the latent period of coronavirus disease 2019 (COVID-19). *Clinical Infectious Diseases*, 74(9):1678–1681, 2022.

Xiaoguang Xu, Theodore Kypraios, and Philip D O'Neill. Bayesian non-parametric inference for stochastic epidemic models using Gaussian processes. *Biostatistics*, 17(4):619–633, 2016.

Qingyu Zhan, Kristina M Babler, Mark E Sharkey, Ayaaz Amirali, Cynthia C Beaver, Melinda M Boone, Samuel Comerford, Daniel Cooper, Elena M Cortizas, Benjamin B Currall, et al. Relationships between SARS-CoV-2 in wastewater and COVID-19 clinical cases and hospitalizations, with and without normalization against indicators of human waste. *Acs Es&T Water*, 2(11):1992–2003, 2022.

Yawen Zhang, Mengsha Cen, Mengjia Hu, Lijun Du, Weiling Hu, John J Kim, and Ning Dai. Prevalence and persistent shedding of fecal SARS-CoV-2 RNA in patients with COVID-19 infection: A systematic review and meta-analysis. *Clinical and Translational Gastroenterology*, 12(4), 2021.

Alessandro Zulli, Annabelle Pan, Stephen M Bart, Forrest W Crawford, Edward H Kaplan, Matthew Cartter, Albert I Ko, Marcela Sanchez, Cade Brown, Duncan Cozens, et al. Predicting daily COVID-19 case rates from SARS-CoV-2 RNA concentrations across a diversity of wastewater catchments. *FEMS microbes*, 2:xtab022, 2021.

Appendix A

Additional Material for Chapter 3

A.1 Methods

A.1.1 SEIR model Used for Simulation

Here we describe in further detail the SEIR model used to simulate the data analyzed in this chapter. The SEIR model describes an infectious disease outbreak of a homogeneously mixing population, with the population divided into four compartments: susceptible, exposed (infected but not yet infectious), infectious, and removed. The SEIR model is represented as a four dimensional continuous time Markov jump process, $\mathbf{G(t)} = (S(t), E(t), I(t), R = (t))$. It can be defined in terms of rate parameters such that

$$
\begin{aligned}
&P(\mathbf{G}(t+dt) = (s-1, e+1, i, r) \mid \mathbf{G}(t) = (s, e, i, r)) = \beta_t \times i \times s/N \times dt + o(dt),\\
&P(\mathbf{G}(t+dt) = (s, e-1, i+1, r) \mid \mathbf{G}(t) = (s, e, i, r)) = \gamma \times e \times dt + o(dt),\\
&P(\mathbf{G}(t+dt) = (s, e, i-1, r+1) \mid \mathbf{G}(t) = (s, e, i, r)) = \nu \times i \times dt + o(dt).
\end{aligned}
$$

We use the well known Gillespie algorithm popularized in [Gillespie, 1977] to simulate from this model, as implemented in the `stemr` R package [Fintzi et al., 2022]. Here γ is the inverse of the mean latent period, and ν is the inverse of the mean infectious period. We describe the infectiousness of the disease through the time-varying transmission rate parameter β_t. With this model, the time-varying basic reproduction number, $R_{0,t}$, and effective reproduction number, R_t, are defined as

$$\begin{aligned} R_{0,t} &= \frac{\beta_t}{\nu}, \\ R_t = R_{0,t} &\times \frac{S(t)}{N}. \end{aligned}$$

By fixing the trajectory of $R_{0,t}$, we fix the trajectory of both β_t but not R_t, because the susceptible population changes stochastically. To simulate case data, we track cumulative incidence through a variable $C(t)$, which counts the transitions from the E to the I state. Cases are then generated at a daily time-scale using the negative binomial model described in the methods section, changing the mean of the model so that, for day t:

$$O_t \mid \mathbf{G}(t), \gamma, \nu, \boldsymbol{\beta}_{0:t}, \rho, \kappa, M_t \sim \text{Neg-Binom}(\rho \times M_t \times (C(t) - C(t-1)), \kappa).$$

A.1.2 True R_t Curves of Scenario 1

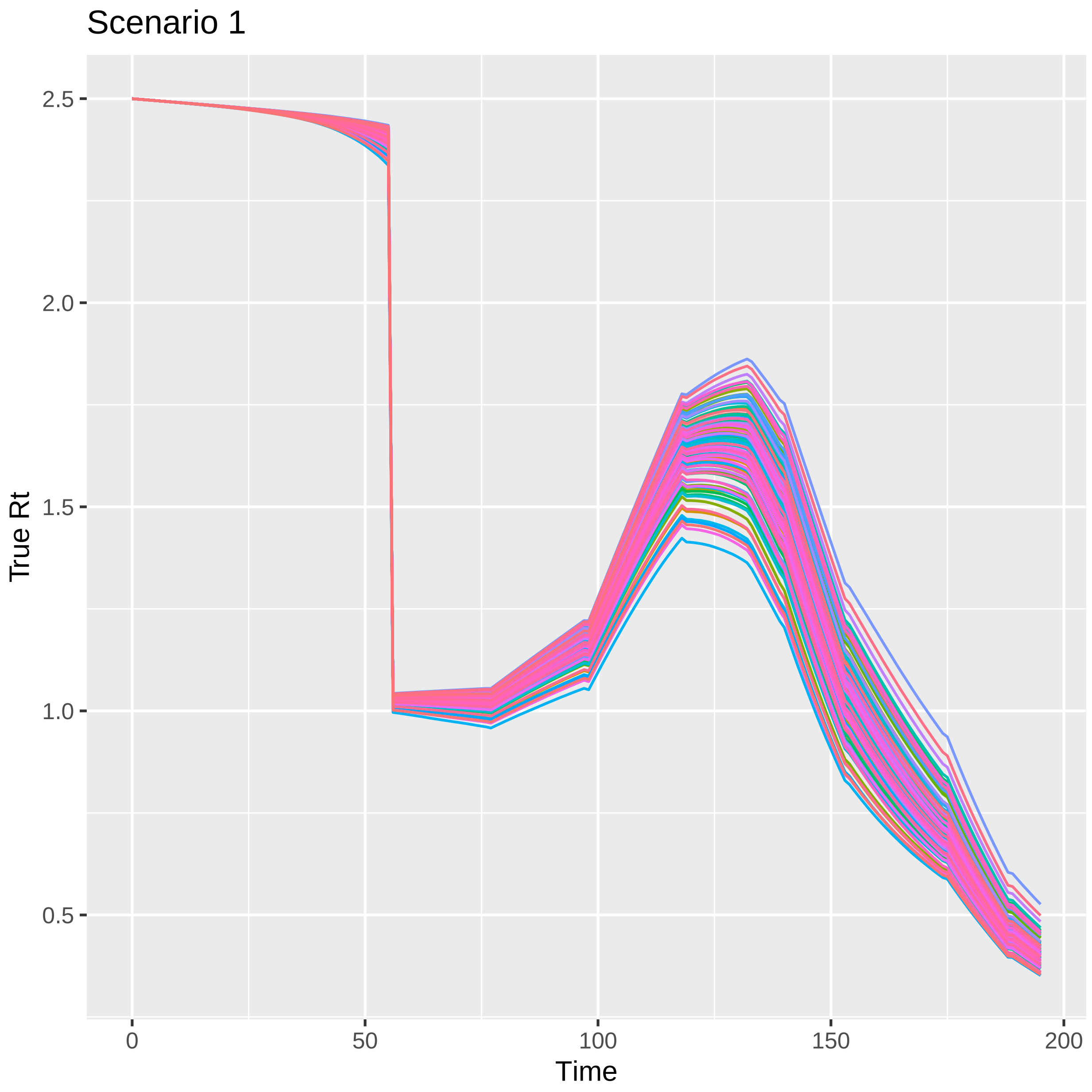


Figure A.1: True R_t curves for Scenario 1. The trajectory of R_0 is fixed, but because the number of susceptibles changes stochastically, each individual realization of the simulation has a slightly different R_t trajectory.

Table A.1: Priors used by the Rt-estim-normal method in the simulation study.

Parameter	Prior	Prior Median (95% Interval)
σ	Truncated-normal(0, 0.1^2)	0.067 (0.0033, 0.26)
λ	Exponential(0.3)	2.31 (0.08, 12.26)
$\log R_1$	Normal(0, 0.2^2)	0.00 (-0.39 0.39)
ψ	Normal(10, 2^2)	10 (6.11, 13.88)
α	Normal(0.02, 0.05^2)	0.02 (-0.08, 0.12)
$\frac{1}{\phi}$	Normal(10, 5^2)	10 (0.23, 19.88)

A.1.3 Rt-estim-normal Model and Parameters

Below is the explicit model structure for the Rt-estim-normal model.

$$
\begin{aligned}
\lambda &\sim \exp(\eta) \qquad \text{Hyperprior for unobserved incidence} \\
I_t &\sim \exp(\lambda) \qquad \text{Prior on unobserved incidence for t= -n, -n-1, \ldots 0} \\
\sigma &\sim \text{Truncated-Normal}(\mu_\sigma, \sigma_\sigma^2) \\
\log R_1 &\sim \text{Normal}(\mu_{r1}, \sigma_{r1}^2) \qquad \text{Prior on } R_1 \\
\log R_t | \log R_{t-1} &\sim \text{Normal}(\log R_{t-1}, \sigma) \qquad \text{Random Walk prior on } R_t \\
\psi &\sim \text{Normal}(\mu_\psi, \sigma_\psi^2) \qquad \text{Prior on variance parameter for incidence} \\
I_t | I_{-n}, \ldots, I_{t-1} &\sim \text{Normal}(R_t \sum_{s<t} I_s g_{t-s}, (R_t \sum_{s<t} I_s g_{t-s} * \psi)^2) \qquad \text{Model for incidence} \\
\alpha &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2) \qquad \text{Prior on case detection rate} \\
y_t &= \alpha_t \sum_{s<t} I_s d_{t-s} \qquad \text{Mean of observed data model} \\
\frac{1}{\phi} &\sim \text{Normal}(\mu_\phi, \sigma_\phi^2) \qquad \text{Prior on dispersion parameter for observed data} \\
Y_t &\sim \text{Neg-Binom}(y_t, \phi) \qquad \text{Observed data model}
\end{aligned}
$$

The same priors were used for all simulations. They are described in Table A.1.

A.1.4 Assessing Model Convergence in Simulations

For Rt-estim-gamma, we assessed the minimum and maximum of the Rhat diagnostic, as well as the minimum and maximum effective sample size for each parameter. We considered maximum values of Rhat below 1.05 to indicate convergence, and considered effective sample size above 100 to be adequate. There were two instances in our original run of all simulations where the diagnostics were above these thresholds. For those specific simulations, we changed the seeds used to change the initial values of the MCMC, which led to convergence.

A.1.5 Discretizing Distributions

The weights g_{t-s} and d_{t-s} used through the paper are discretized versions of continuous probability distributions. The number of discretized values to create was usually set to be the number of observed data points (occasionally with one additional value). For each value u greater than 1, the discretized value was calculated as

$$g_u = F(u + 0.5) - F(u - 0.5), u = 2, \ldots$$

where $F(u)$ is the cumulative distribution function for the distribution being discretized. For $u = 1$, in the case of the generation time distribution we used

$$g_1 = F(1.5),$$

but for the latent distribution, we used

$$g_0 = F(0.5)$$

$$g_1 = F(1.5) - F(0.5)$$

in order to have a discretized value corresponding to 0.

A.1.6 Using Rt-estim-gamma with Real Data

In practice, we have found it often necessary to provide reasonable initial values to start the Hamiltonian-Monte Carlo algorithm when applying the Rt-estim-gamma model to real data. Even so, running 4 chains, only three converged. We ran chains for 6000 iterations, discarding half as burn-in, and kept results only when the maximum Rhat value was calculated to be less than 1.05, with minimum bulk ESS and tail ESS above 100 as calculated using `rstan`. In all counties, the tail and bulk ESS for the estimates for the effective reproduction number had a minimum value of 1000. For the effective reproduction number, we first used `EpiEstim` to estimate the effective reproduction number, then used the median estimate from the posterior as the starting point for the effective reproduction number in Rt-estim-gamma. For incidence, we used the median of the overall case detection rate prior times the observed cases for the corresponding day. For all other model parameters, we used the mean of the prior distribution as the starting point.

A.1.7 Creating Generation Time Distributions for Delta and Omicron Variants

We created generation time distributions for these variants by searching for parameters such that the new distributions had the appropriate new mean generation time, while preserving the standard deviation of the original distribution. We used a squared error loss function as a cost function, using the squared difference in a candidate distribution's mean vs the desired mean plus the squared difference in the candidate distribution's standard deviation vs the desired standard deviation. The estimates of the candidate distribution's mean and standard deviations were method of moment estimates from 100,000 samples generated in R.

A.2 Results

A.2.1 Posterior Predictive Distribution for Three Scenarios

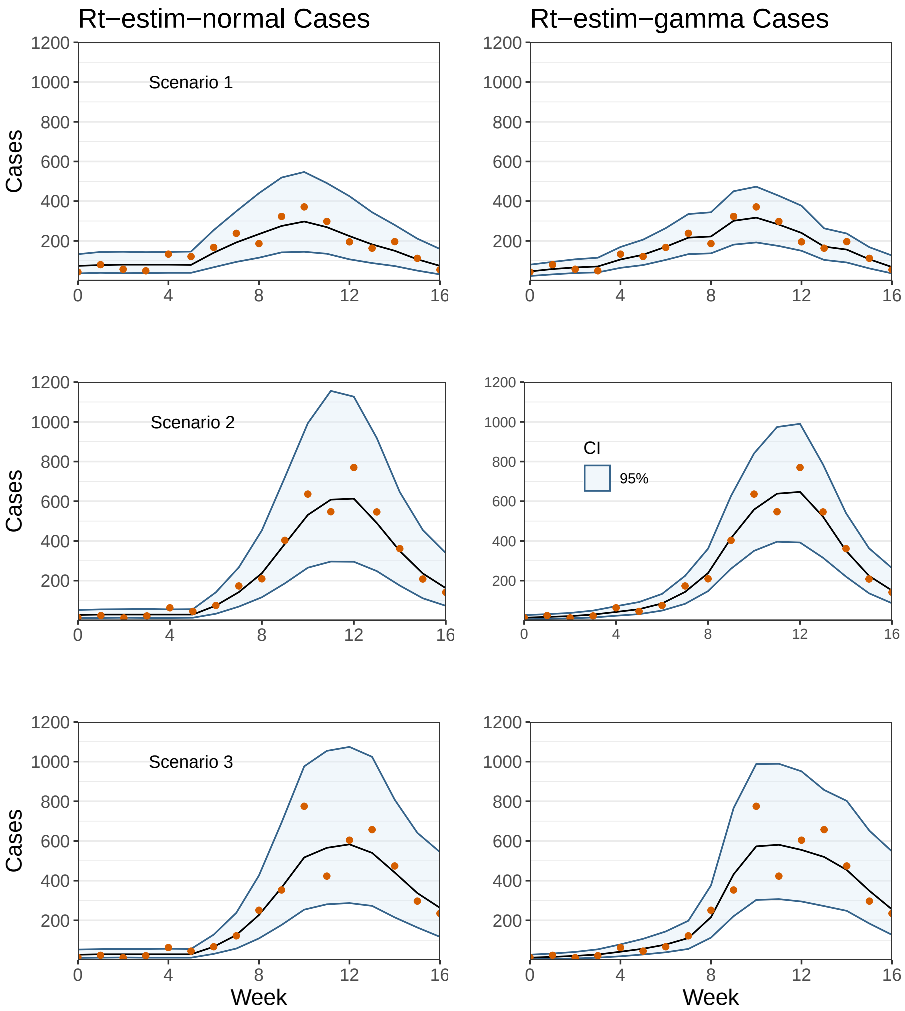


Figure A.2: Posterior predictive distributions for reported cases using two different R_t estimation methods for three simulated data sets under different testing scenarios. True incidence trajectories are colored in red, black lines represent median estimates from the posterior distribution, blue shaded ares are 95% credible intervals.

A.2.2 Example Negative-Binomial Spline Posterior Predictive

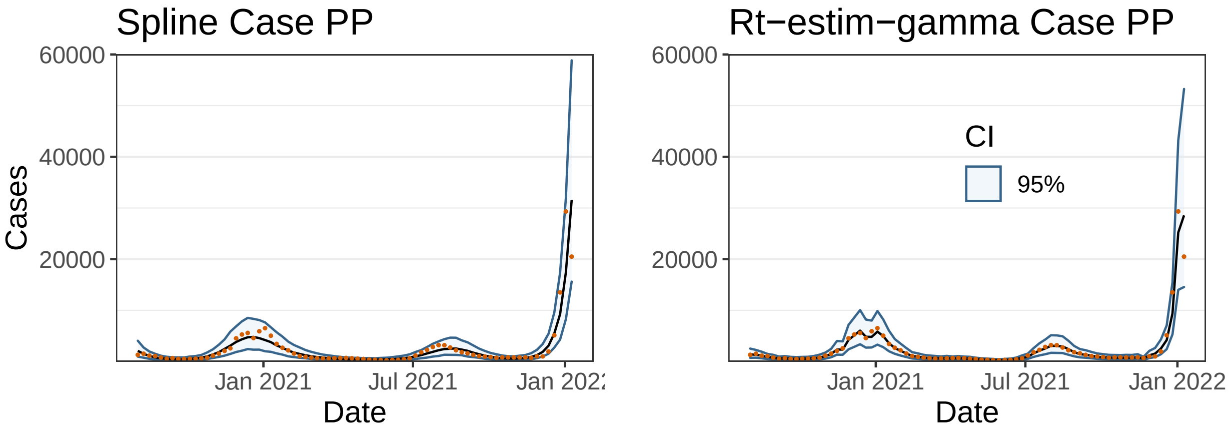


Figure A.3: Posterior predictive plots of cases of SARS-CoV-2 in Alameda County, CA from August 20th 2020 through January 9th 2022. The left plot shows the posterior predictive distribution for a negative binomial spline, while the right plot shows the posterior predictive from the Rt-estim-gamma model. Black lines represent medians, blue bands 95% credible intervals, and dots are observed cases.

A.2.3 Secondary Simulation Frequentist Metrics

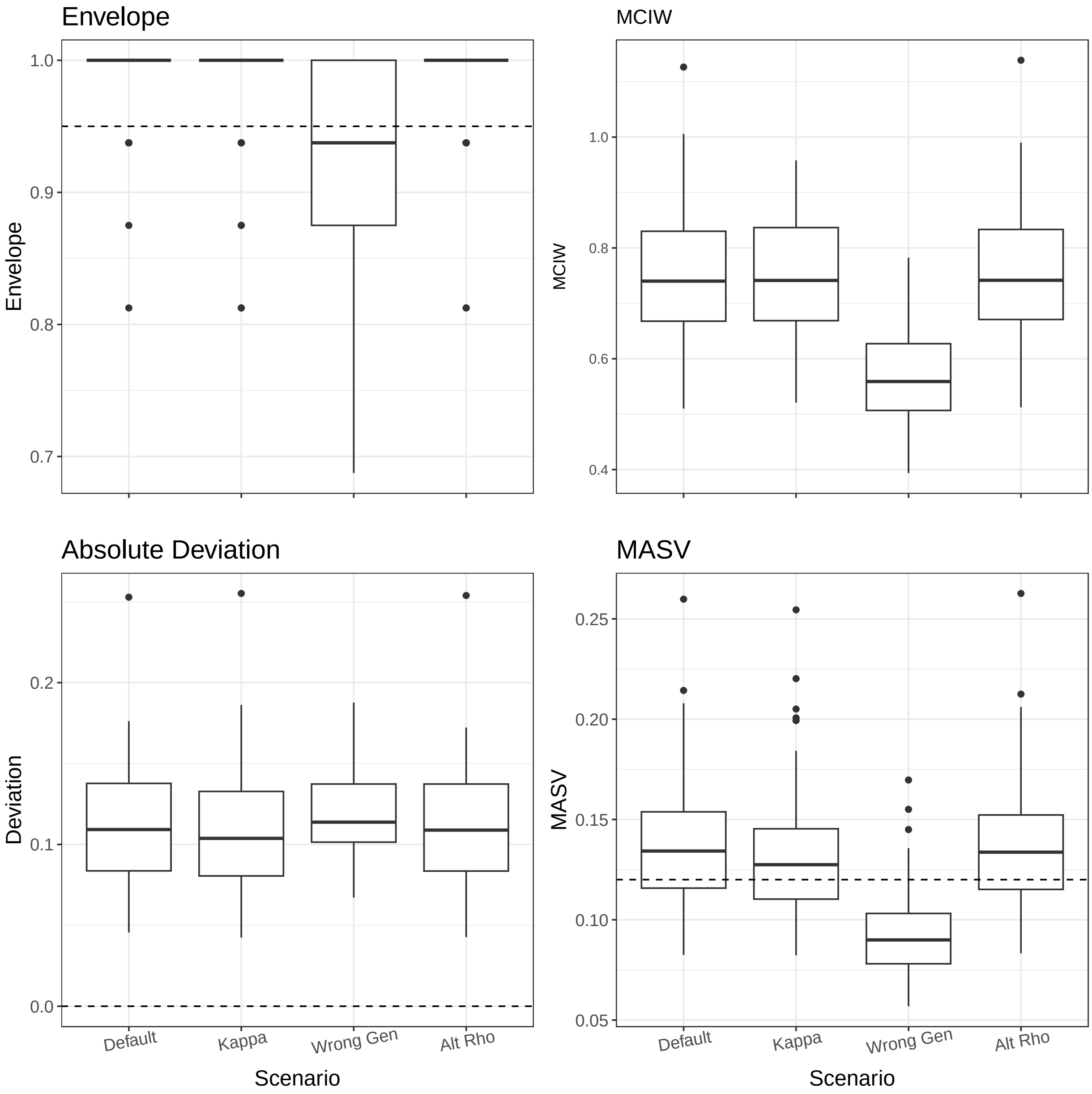


Figure A.4: Frequentist metrics for Rt-estim-gamma applied to simulated epidemics from Scenario 3 with alternative model parameters. Default refers to the Scenario 3 results reported in the main text. Kappa refers to the choosing the priors for the overdispersion parameter of the case observation model by using a spline on the actual data being analyzed. Wrong gen refers to using a generation time distribution with rate parameters twice as large as the correct rate parameters. Alt rho refers to using a prior for ρ based on the 25% quantile of tests, rather than the 50% quantile used in the main analysis. The dashed lines in the bottom row represent the true absolute standard deviation. Middle lines are medians across 100 simulations, hinges are upper and lower quartiles, whiskers are at most 1.5 times the interquartile range from the median.

A.2.4 Using `EpiEstim` to Estimate the Effective Reproduction Number in CA

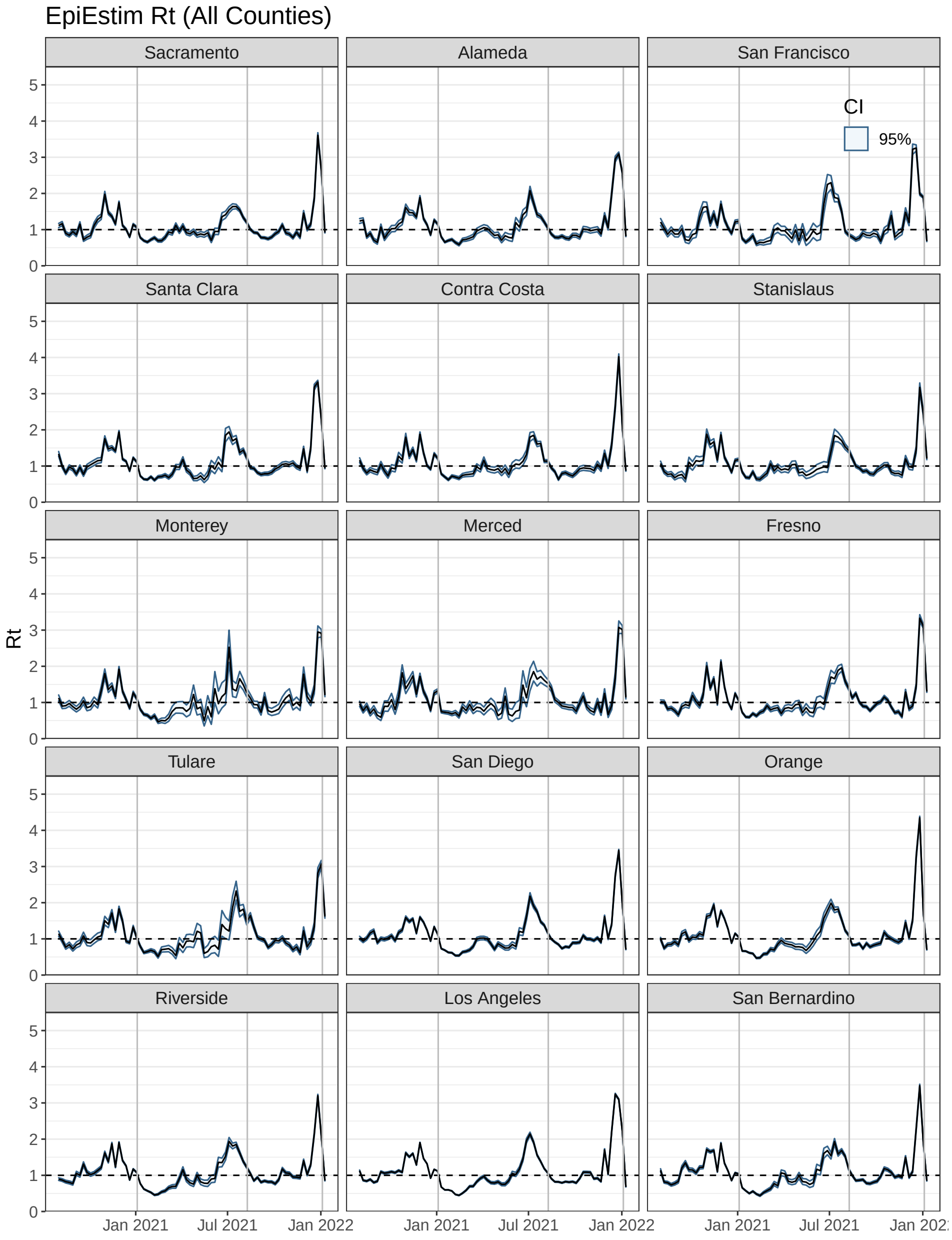


Figure A.5: Estimates of the effective reproduction number in fifteen counties of California from August 2nd 2020 through January 9th 2022 using `EpiEstim`. Black lines represent medians, blue bars are 95% credible intervals.

A.2.5 Incidence Posterior and Case Posterior Predictive Plots from Rt-estim-gamma for Fifteen California Counties

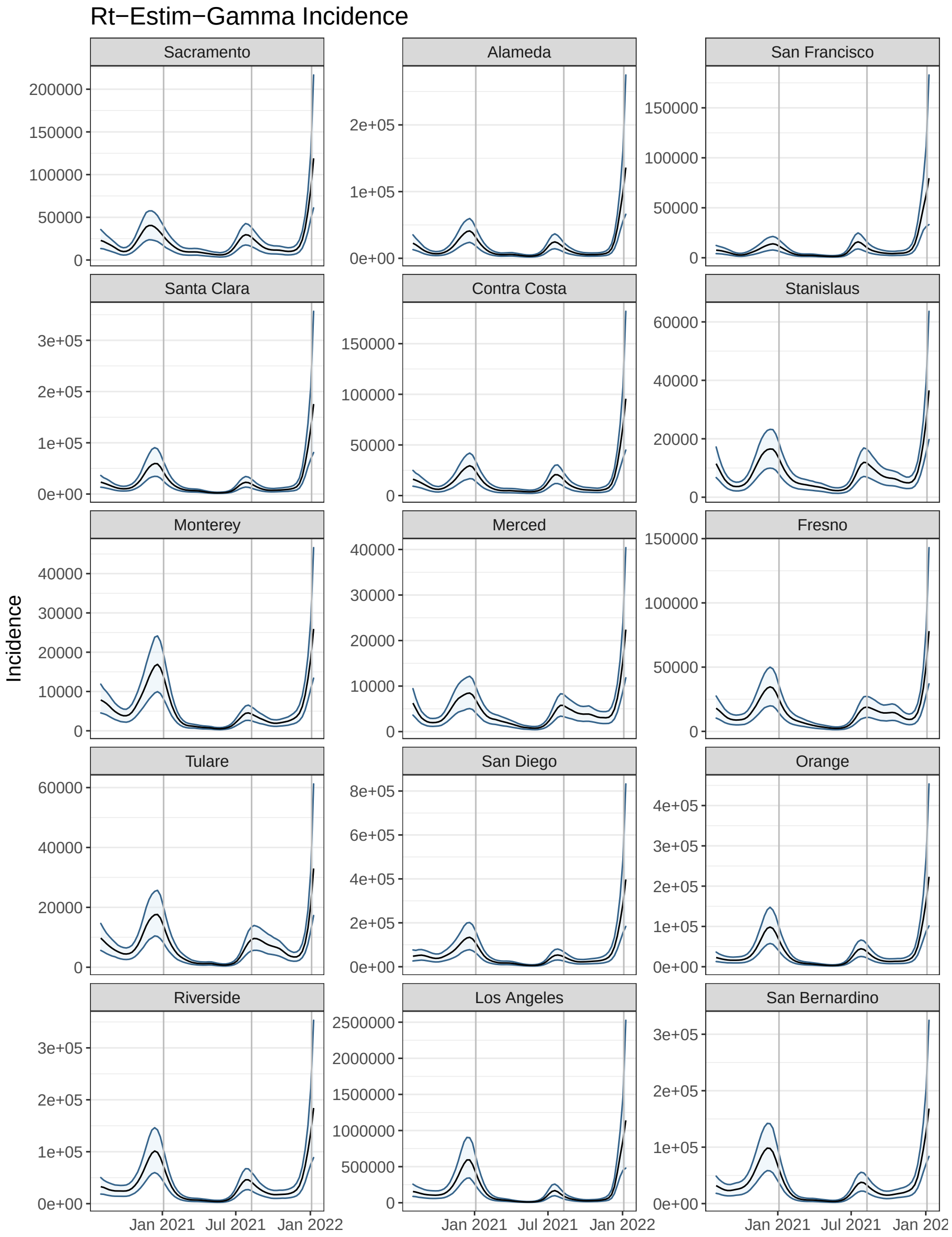


Figure A.6: Estimates of incidence of SARS-CoV-2 from Rt-estim-gamma applied to fifteen counties in California, USA from August 2nd 2020 through January 15th 2022. Blue shaded regions are 95% posterior credible intervals. Black lines are medians. Grey vertical lines mark the maximum statewide cases reported for the original winter 2020 wave, the delta-variant wave, and the omicron-variant wave.

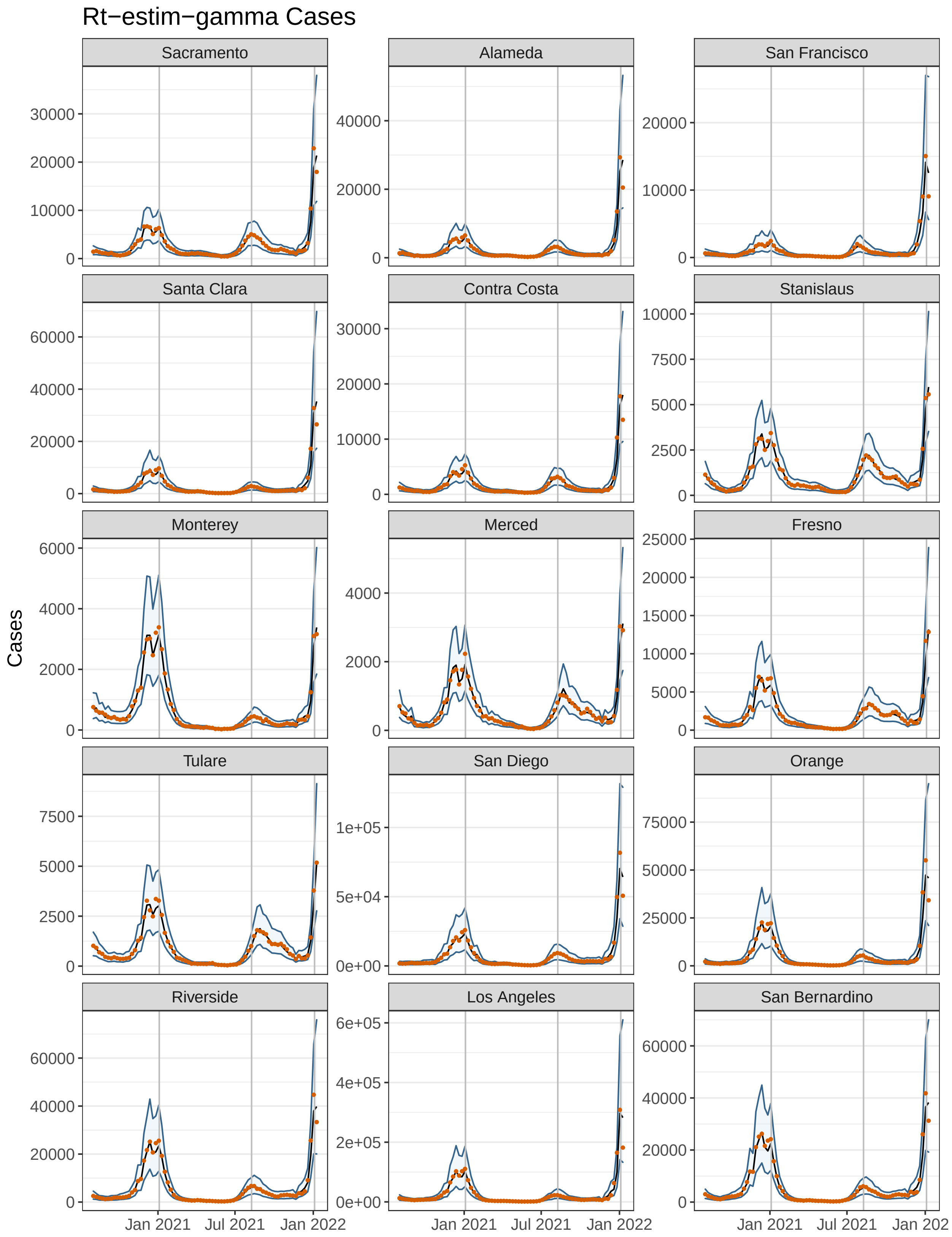


Figure A.7: Posterior predictive estimates of reported cases of SARS-CoV-2 from Rt-estim-gamma applied to fifteen counties in California, USA from August 2nd 2020 through January 15th 2022. Blue shaded regions are 95% posterior credible intervals. Black lines are medians. Grey vertical lines mark the maximum statewide cases reported for the original winter 2020 wave, the delta-variant wave, and the omicron-variant wave. Red dots are observed case counts.

A.2.6 Comparing Rt-estim-normal and Rt-estim-gamma Applied to Fifteen California Counties

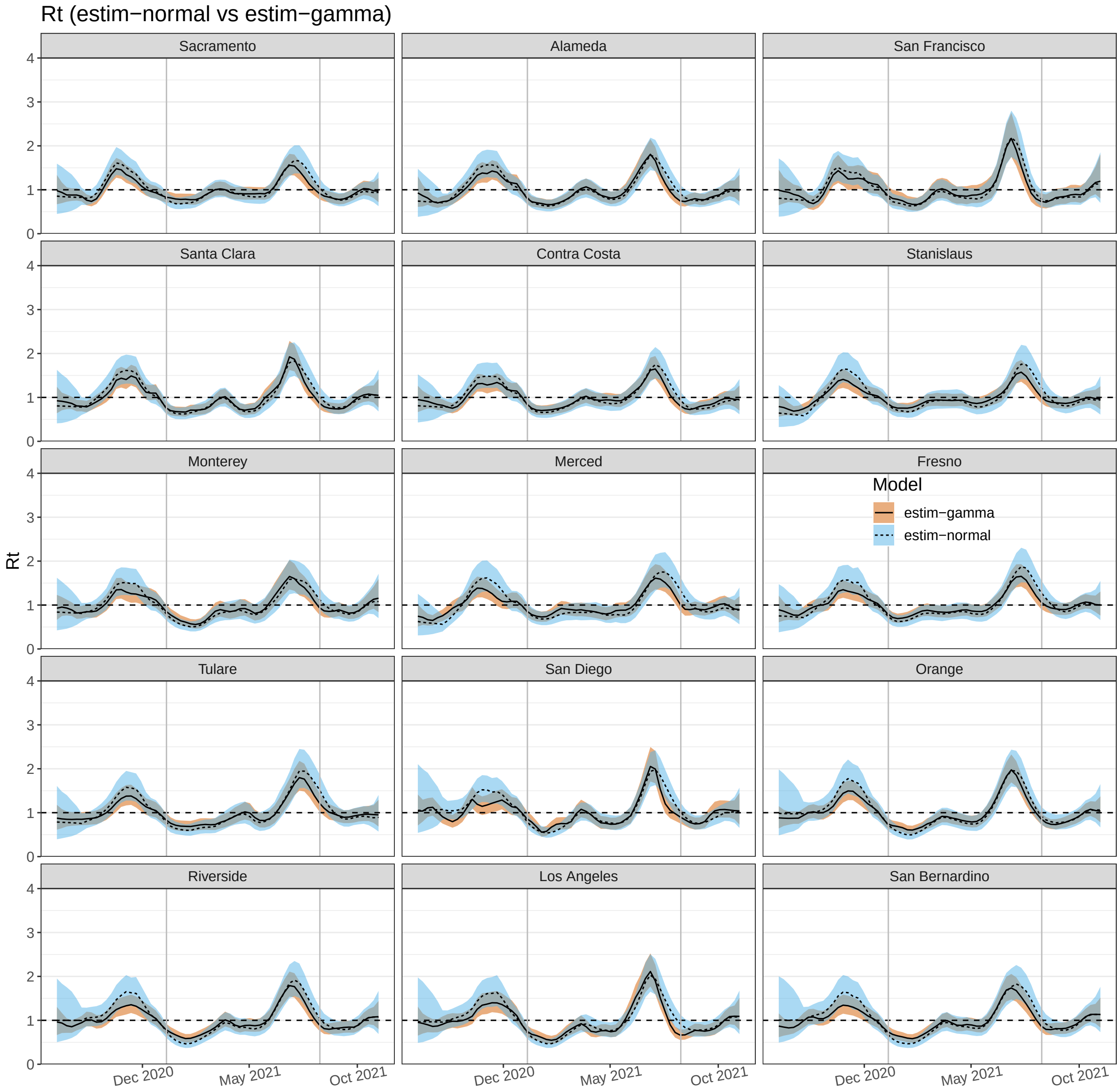


Figure A.8: Estimates of the effective reproduction number of SARS-CoV-2 from Rt-estim-gamma and Rt-estim-normal applied to fifteen counties in California, USA from August 2nd 2020 through November 6th 2021. Blue and brown shaded regions are 95% posterior credible intervals. Black and dotted lines are medians. Grey vertical lines mark the maximum statewide cases reported for the original winter 2020 wave and the delta-variant wave. Blue shading and black lines come from estimates using Rt-estim-normal as opposed to brown shading with dotted lines, which denote estimates using Rt-estim-gamma.

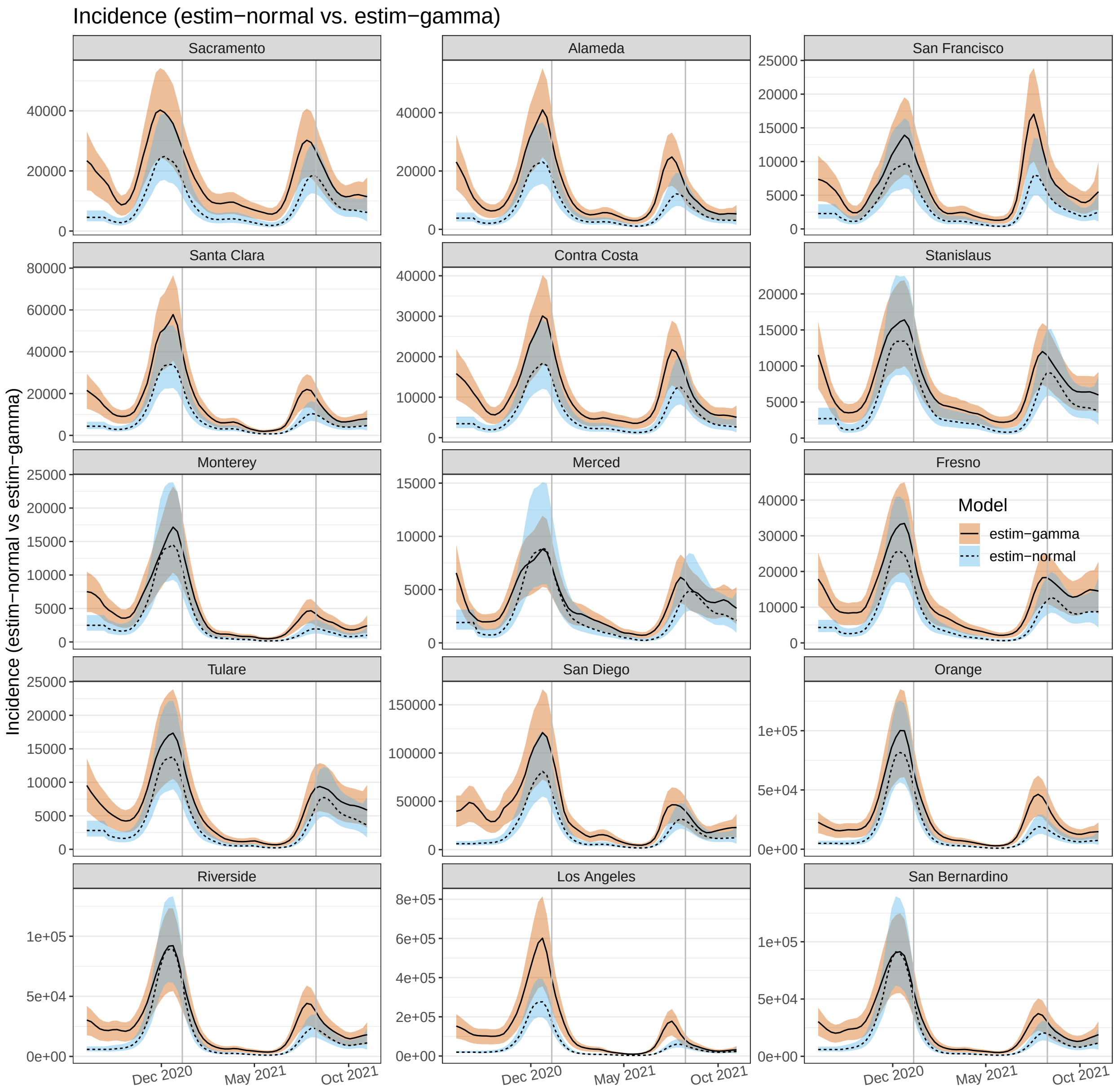


Figure A.9: Estimates of incidence of SARS-CoV-2 from Rt-estim-gamma and Rt-estim-normal applied to fifteen counties in California, USA from August 2nd 2020 through November 6th 2021. Blue and brown shaded regions are 95% posterior credible intervals. Black and dotted lines are medians. Grey vertical lines mark the maximum statewide cases reported for the original winter 2020 wave and the delta-variant wave. Blue shading and black lines come from estimates using Rt-estim-normal as opposed to brown shading with dotted lines, which denote estimates using Rt-estim-gamma.

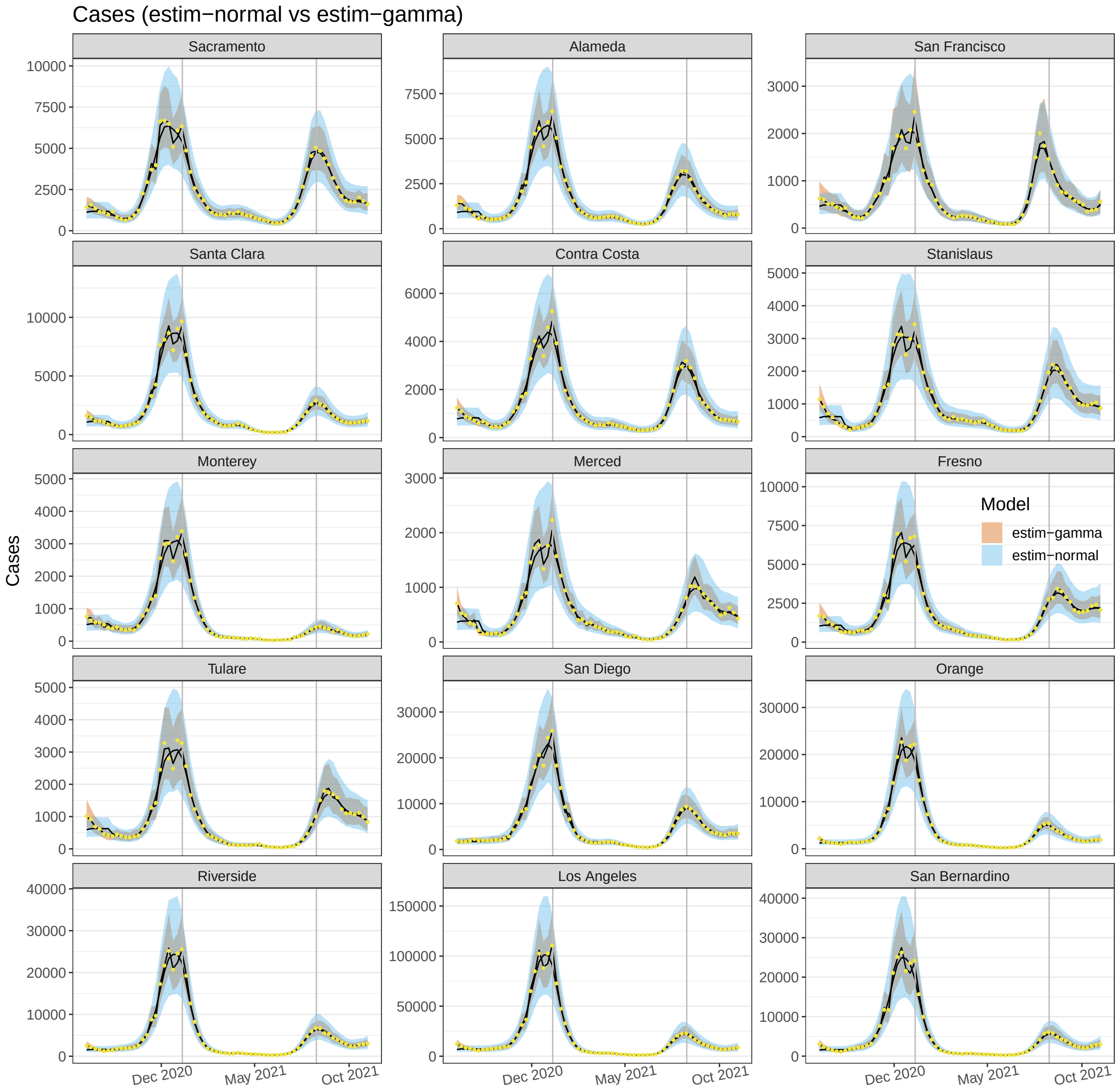
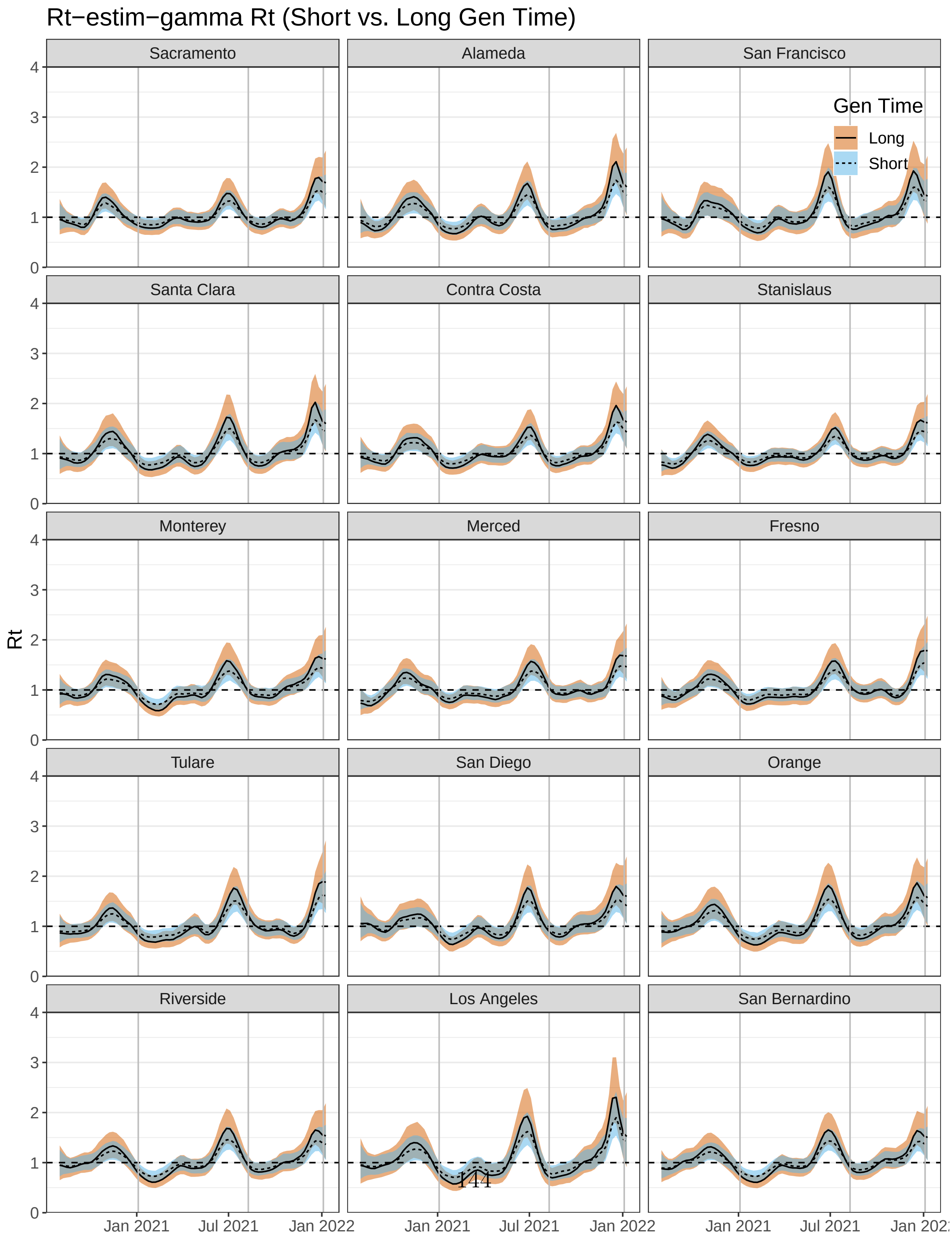


Figure A.10: Estimates of observed cases of SARS-CoV-2 from Rt-estim-gamma and Rt-estim-normal applied to fifteen counties in California, USA from August 2nd 2020 through November 6th 2021. Blue and brown shaded regions are 95% posterior predictive intervals intervals. Black and dotted lines are medians. Grey vertical lines mark the maximum statewide cases reported for the original winter 2020 wave and the delta-variant wave. Blue shading and black lines come from estimates using Rt-estim-normal as opposed to brown shading with dotted lines, which denote estimates using Rt-estim-gamma. Yellow dots are observed cases

A.2.7 Comparing Rt-estim-gamma Applied to Fifteen California Counties Using a Mean Generation Time of 5.5 Days vs 9.7 Days



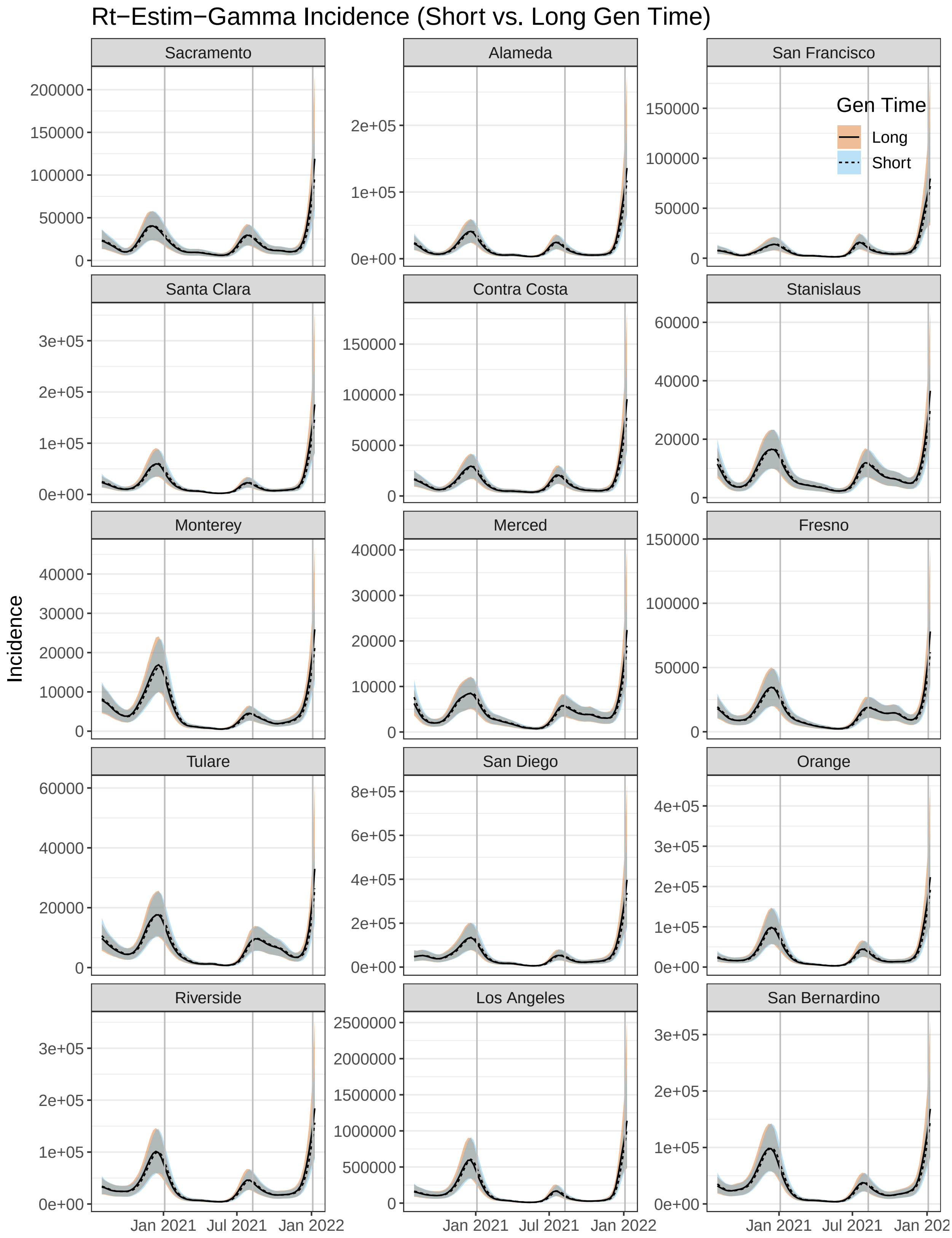


Figure A.12: Estimates of incidence of SARS-CoV-2 from Rt-estim-gamma applied to fifteen counties in California, USA from August 2nd 2020 through January 15th 2022. Blue and brown shaded regions are 95% posterior credible intervals. Black and dotted lines are medians. Grey vertical lines mark the maximum statewide cases reported for the original winter 2020 wave, the delta-variant wave, and the omicron-variant wave. Blue shading and dotted lines come from estimates using mean generation time of 5.5 days as opposed to

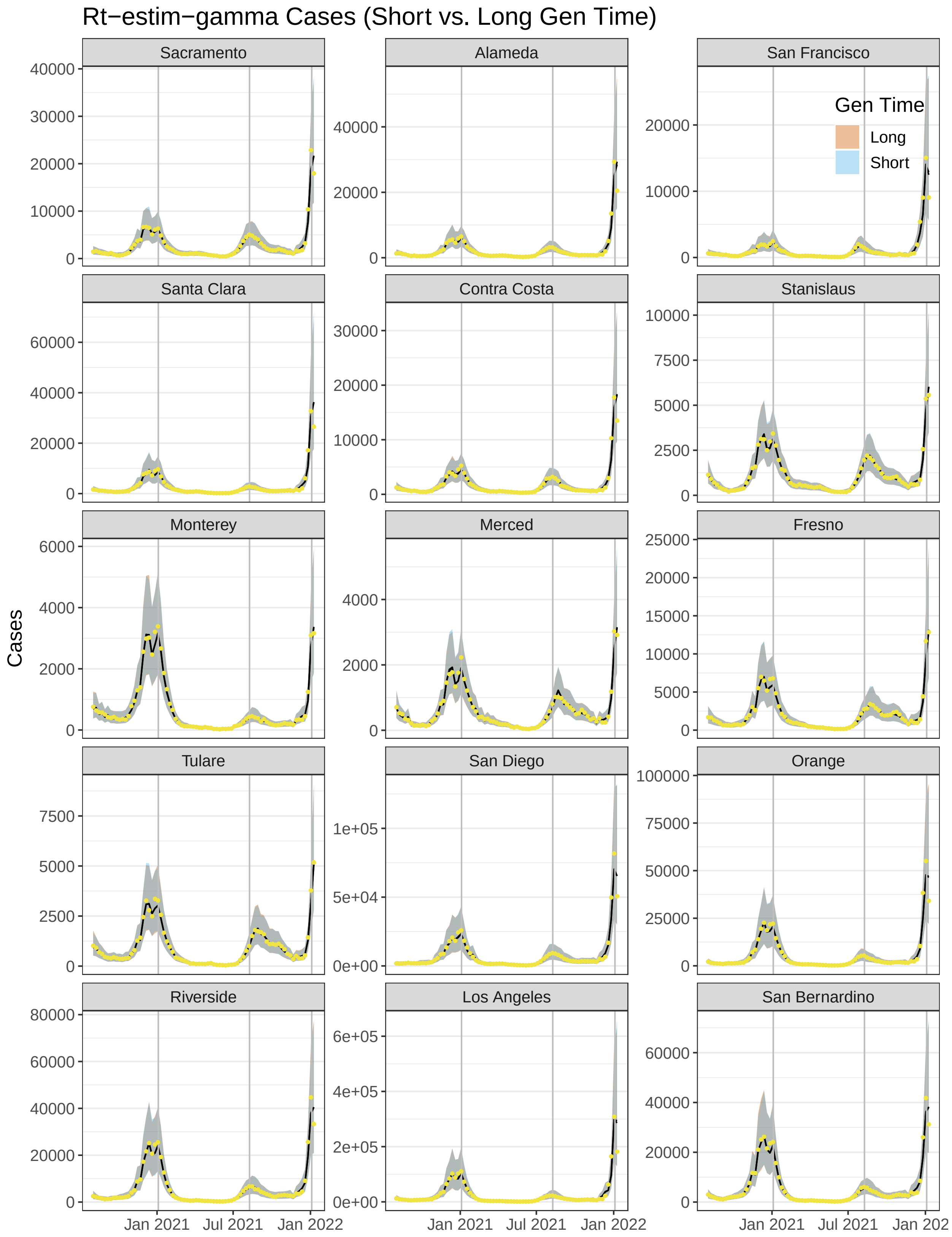


Figure A.13: Posterior predictive estimates of reported cases of SARS-CoV-2 from Rt-estim-gamma applied to fifteen counties in California, USA from August 2nd 2020 through January 15th 2022. Blue and brown shaded regions are 95% posterior credible intervals, grey represents overlap between the two estimates. Grey vertical lines mark the maximum statewide cases reported for the original winter 2020 wave, the delta-variant wave, and the omicron-variant wave. Yellow dots are observed case counts. Blue shading come from estimates using

A.2.8 Prior and Posterior of Fixed Parameters

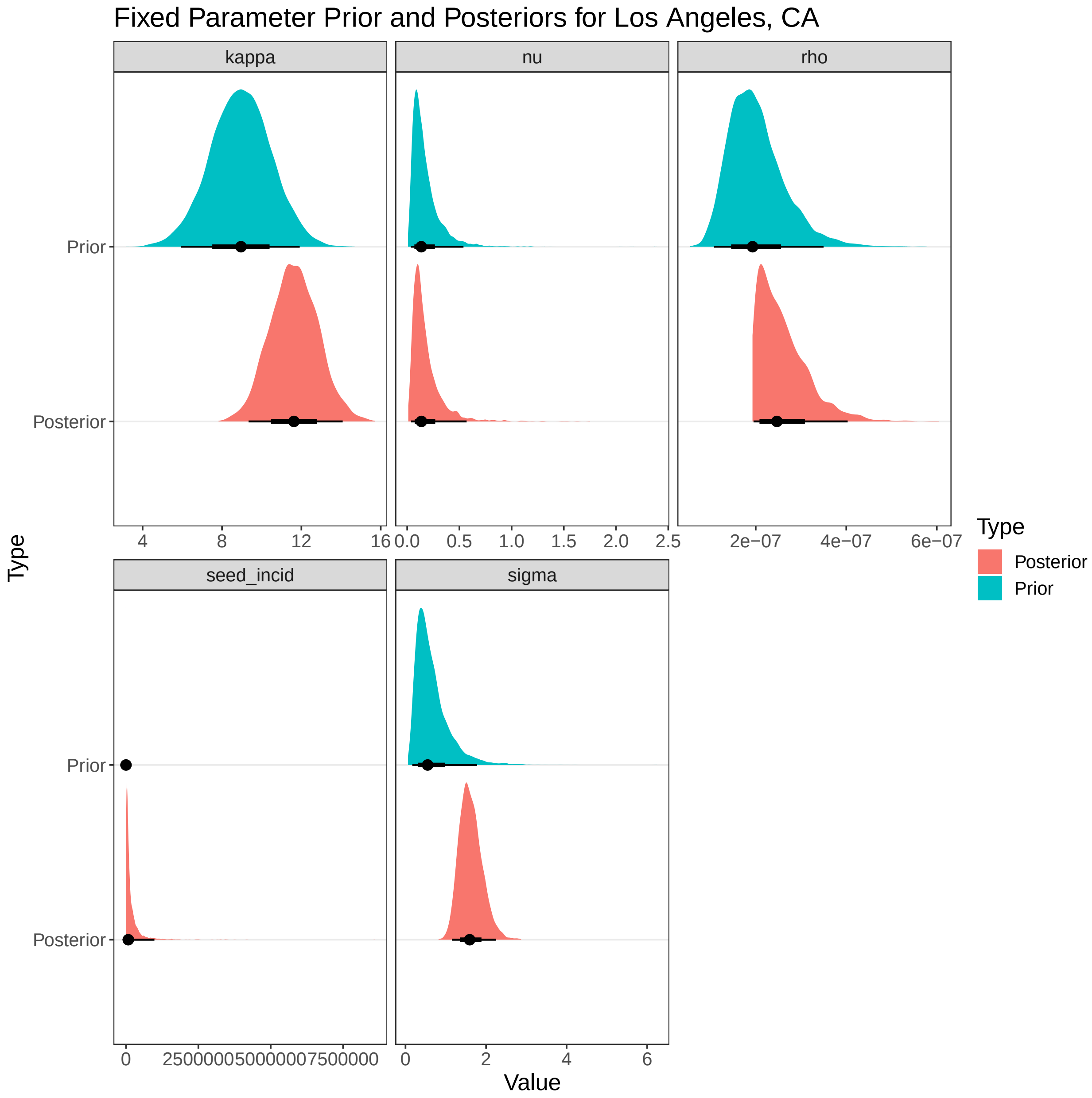


Figure A.14: Priors and posteriors for fixed parameters from Rt-estim-gamma fit to Los Angeles, CA data using the Sender generation time. The seed_incid refers to the first unobserved incidence used by the model.

Appendix B

Additional Material for Chapter 4

B.1 Methods

B.1.1 Choosing a λ Prior for SARS-CoV-2 Using a Stochastic SEIRR

The parameter λ controls how much pathogen genomic concentrations are attributed to the total number of infectious individuals, versus the total number of recently recovered individuals. Previous studies on the timing and magnitude of shedding SARS-CoV-2 RNA have necessarily been concerned with the shedding dynamics of individuals [Benefield et al., 2020, Miura et al., 2021, Hoffmann and Alsing, 2023]. However, our model is concerned with the shedding dynamics of populations. We wanted to create a prior for λ that incorporates what we know about shedding dynamics for SARS-CoV-2, but needed a way to translate that information into an appropriate prior for population level shedding dynamics. Furthermore, our model assumes the relationship between compartment counts and gene concentrations is linear, we wanted to make sure our process for constructing the prior for λ incorporated this linearity assumption. Thus, we used a simulation study using an individual-level engine that could incorporate our prior individual level information combined with linear regression models in order to elicit an appropriate prior for λ. We simulated 1000 epidemics from an agent-based stochastic SEIRR model (the stochastic equivalent to an SEIR model with two R compartments, see Web Section A.2.1), where the times of each individual's transitions between different model states were recorded. The population was 1000, R_0 was set to 2, there were 5 initially infectious agents, and all other agents were susceptible. For individual i in the I or $R1$ state infected at time t_i, the concentration of pathogen genomes associated with their shedding at time l was modeled as a random variable $Z_i(t_i, l)$, where

$$Z_i(t_i, l) \sim 10^{\text{Normal}(\mu_i(l-t_i), 1.09)}$$

and the value of $\mu_i(l - t_i)$ was calculated using the consensus shedding load profile of SARS-CoV-2 pathogen RNA generated in [Nourbakhsh et al., 2022] by synthesizing previous studies [Benefield et al., 2020, Miura et al., 2021, Hoffmann and Alsing, 2023].

The value 1.09 is the average variation in genome concentrations on the log base 10 scale amongst individuals over the course of an infection. We calculated this global average variation in concentrations by averaging over the empirical standard deviations of SARS-CoV-2 RNA shed by individuals 6 to 22 days after symptom onset, using the data available from [Hoffmann and Alsing, 2023], which uses data collected by Wölfel et al. [2020], Han et al. [2020], and Lui et al. [2020]. The population level concentration of genomes at time l was the sum of all the individual genome concentrations in the I and $R1$ states at time l divided by the total population size.

To account for uncertainty in the parameters governing how long individuals spend in the I and $R1$ compartments, for each simulation, γ, ν and η were chosen from the priors we used when fitting the EIRR model to SARS-CoV-2 data (Table A2).

Then, for each simulation, we fit a linear model (constrained to positive coefficients using a method described in [Friedman et al., 2010]):

$$E[\text{Total genome concentration}] = \beta_1 \times \text{Prevalence in } I + \beta_2 \times \text{Prevalence in } \mathit{R1}$$

and calculated $\lambda = \beta_1/(\beta_1 + \beta_2)$. Finally, we constructed a logit-normal prior for λ that matches the 95% quantiles of these 1000 λ values by minimizing the squared error of the logit-normal prior quantiles and the 95% quantiles from our 1000 λ values using the Nelder-Mead algorithm implemented in the `optim` function in `R` [R Core Team, 2020]. Further details of the simulation protocol are available in Web Sections A.2.1, A.2.2, and A.2.3.

B.1.2 Details of the Huisman Method

We compare the EIRR-ww model to the Huisman et al. [2022b] method. This method is a variation on the well known `EpiEstim` method [Cori et al., 2013]. Pathogen genome concentrations are modeled as a function of incidence (newly infected individuals) counts via a convolution equation:

$$C_i = M \sum_j w_{i-j} I_j. \tag{B.1}$$

Here C_i is the concentration at time i, I_j is the number of new infections in the period $(j-1, j]$, and w_{i-j} is a weight derived from discretizing the assumed individual shedding load profile describing how many pathogen genomes an infected individual sheds over time. M is a constant value translating counts of individuals to counts of pathogens, the Huisman method assumes M is the lowest observed concentration in the data set. A time series of incidence is constructed via a deconvolution algorithm, and then used as the inputs into `EpiEstim`. `EpiEstim` is a method inspired by branching process approximations of the spread of infectious disease where infectious individuals generate new infectious individuals in a Crump-Mode-Jager process [Fraser, 2007, Cori et al., 2013, Pakkanen et al., 2023]. The core concept is to model current incidence as a function of previous incidence and the effective reproduction number through the so-called Renewal Equation:

$$\mathrm{E}(I_t \mid \mathbf{I}_{1:t}, R_t) = R_t \sum_{u=1}^{t-1} I_u g_{t-u}. \tag{B.2}$$

Here g_{t-u} are values from the discretized generation time distribution, the distribution of the time between one person becoming infected, and subsequently infecting someone else. `EpiEstim` models incidence conditioned on previous incidence and the effective reproduction number as a Poisson random variable, and holds R_t constant for a window of time, creating a smooth estimate by repeatedly re-estimating R_t for all such windows in the time series. Huisman et al. [2022b] repeats this pipeline multiple times via a bootstrap method to generate uncertain estimates of the effective reproduction number.

B.1.3 Priors for Models with the S Compartment

We will assume the basic reproduction number is constant in a time interval $(k_i, k_{i+1}]$, defining it as $R_{0,k_i} = \frac{\beta_{k_i}}{\nu}$. Let M be the total number of time intervals of interest. Let $\mathbf{R_0} = (R_{0,0}, R_{0,k_1}, \ldots, R_{0,k_M})$, be the vector of basic reproduction numbers. We use a random walk prior so that

$$\begin{aligned}
R_{0,0} &\sim \text{Log-Normal}(\mu_{0,0}, \sigma_{0,0}),\\
\sigma &\sim \text{Log-Normal}(\mu_{rw}, \sigma_{rw}),\\
\log(R_{0,k_{i+1}}) | \log(R_{0,k_i}), \sigma &\sim \text{Normal}(\log(R_{0,k_i}), \sigma_{rw}).
\end{aligned}$$

For this chapter, we assume the basic reproduction number changes on a weekly basis, but it could change according to other time scales. Our model also requires initial conditions in order to solve the system of ODEs. Let N be the population size (assumed to be known). Let P be the population not in the $R2$ compartment at the time the model is fit. For simulations, this value is also known, in real world settings, we assume $N = P$, that is, the difference between the two is negligible in large populations. In the case of the SEIRR-ww model, the initial conditions are calculated as:

$$\begin{aligned}
S(0) &= P * S\_SEIR1, \\
I(0) &= (P - S(0)) * I\_EIR1, \\
R1(0) &= (P - S(0) - I(0)) * R1\_ER1, \\
E(0) &= (P - S(0) - I(0) - R1(0)), \\
R2(0) &= 1,
\end{aligned}$$

where we define S_SEIR1 as the proportion of the population in the S compartment at time 0, I_EIR1 as the proportion of those in the E, I, or $R1$ compartments in the I compartment and $R1\_ER1$ as the proportion of those in the E or $R1$ compartments in the $R1$ compartment. We use logit-normal priors for the proportions. We use a similar technique when using the SEIR-cases model, with one less parameter as there is only one R compartment.

B.1.4 Closed Form Solutions of the EIR/EIRR Models

The systems of ordinary differential equations for the EIR/EIRR models are linear, and thus can be solved in closed form. We used `Mathematica` Version (13.1) to calculate the closed form solution.

B.1.5 Closed Form Solution for the EIRR Model

Define the EIRR model as:

$$\begin{aligned}
\frac{dE}{dt} &= \alpha_t * I - \gamma E \\
\frac{dI}{dt} &= \gamma E - \nu I \\
\frac{dR1}{dt} &= \nu I - \eta R1 \\
\frac{dR2}{dt} &= \eta R1
\end{aligned}$$

Let V be

$$V = \begin{bmatrix} -\gamma & \alpha_t & 0 & 0 \\ \gamma & -\nu & 0 & 0 \\ 0 & \nu & -\eta & 0 \\ 0 & 0 & \eta & 0 \end{bmatrix}.$$

The matrix exponential of V for a fixed α_t is reported on the next page. For initial conditions $M(t_0)$, the solution to the system of ODEs is

$$e^{V(t-t_0)}M(t_0).$$

$a = 4\alpha\gamma + \gamma^2 - 2\gamma\nu + \nu^2$

$b = 3\eta + 3\gamma + 3\nu$

$c = -2\alpha\gamma + 2\eta\gamma + 2\eta\nu + 2\gamma\nu$

$$A = \frac{\left(\frac{1}{2}(\alpha\gamma-\gamma(\eta-\gamma))(-\sqrt{a}-\gamma+\nu)+\gamma(\alpha(\eta-\nu)-\alpha\gamma)\right)e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(-\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu} + \frac{\left(\frac{1}{2}(\alpha\gamma-\gamma(\eta-\gamma))(\sqrt{a}-\gamma+\nu)+\gamma(\alpha(\eta-\nu)-\alpha\gamma)\right)e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu}$$

$$B = \frac{\left(\frac{1}{2}(-\sqrt{a}+\gamma-\nu)(\alpha(\eta-\nu)-\alpha\gamma)+\alpha(\alpha\gamma-\gamma(\eta-\gamma))\right)e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(-\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu} + \frac{\left(\frac{1}{2}(\sqrt{a}+\gamma-\nu)(\alpha(\eta-\nu)-\alpha\gamma)+\alpha(\alpha\gamma-\gamma(\eta-\gamma))\right)e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu}$$

$$C = \frac{\left(\frac{1}{2}(-\sqrt{a}-\gamma+\nu)(\gamma(\eta-\gamma)-\gamma\nu)+\gamma(\alpha\gamma-\nu(\eta-\nu))\right)e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(-\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu} + \frac{\left(\frac{1}{2}(\sqrt{a}-\gamma+\nu)(\gamma(\eta-\gamma)-\gamma\nu)+\gamma(\alpha\gamma-\nu(\eta-\nu))\right)e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu}$$

$$D = \frac{\left(\frac{1}{2}(-\sqrt{a}+\gamma-\nu)(\alpha\gamma-\nu(\eta-\nu))+\alpha(\gamma(\eta-\gamma)-\gamma\nu)\right)e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(-\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu} + \frac{\left(\frac{1}{2}(\sqrt{a}+\gamma-\nu)(\alpha\gamma-\nu(\eta-\nu))+\alpha(\gamma(\eta-\gamma)-\gamma\nu)\right)e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu}$$

$$E = \frac{\gamma\nu e^{-\eta t}}{-\alpha\gamma+\eta^2-\eta\gamma-\eta\nu+\gamma\nu} + \frac{\left(-\gamma\nu\sqrt{a}+\gamma^2(-\nu)-\gamma\nu^2\right)e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{2\left(\frac{1}{4}\left(-\sqrt{4\alpha\gamma+\gamma^2-2\gamma\nu+\nu^2}-\gamma-\nu\right)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu\right)} +$$
$$\frac{\left(\gamma\nu\sqrt{a}+\gamma^2(-\nu)-\gamma\nu^2\right)e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{2\left(\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu\right)}$$

$$F = -\frac{e^{-\eta t}\left(\eta^2\nu-\eta\gamma\nu\right)}{\eta\left(-\alpha\gamma+\eta^2-\eta\gamma-\eta\nu+\gamma\nu\right)} + \frac{\left(\nu^2\sqrt{a}+2\alpha\gamma\nu-\gamma\nu^2+\nu^3\right)e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{2\left(\frac{1}{4}(-\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu\right)} + \frac{\left(-\nu^2\sqrt{a}+2\alpha\gamma\nu-\gamma\nu^2+\nu^3\right)e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{2\left(\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu\right)}$$

$$G = \frac{e^{-\eta t}(-\sqrt{a}+2\eta-\gamma-\nu)(\sqrt{a}+2\eta-\gamma-\nu)}{4\left(-\alpha\gamma+\eta^2-\eta\gamma-\eta\nu+\gamma\nu\right)}$$

$$H = -\frac{\gamma\nu e^{-\eta t}}{-\alpha\gamma+\eta^2-\eta\gamma-\eta\nu+\gamma\nu} + \frac{\eta\gamma\nu e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(-\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu} + \frac{\eta\gamma\nu e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu} +$$
$$\frac{\eta\gamma\nu}{\eta\gamma\nu-\alpha\eta\gamma}$$

$$I = \frac{\nu(\eta-\gamma)e^{-\eta t}}{-\alpha\gamma+\eta^2-\eta\gamma-\eta\nu+\gamma\nu} - \frac{\eta\nu(\sqrt{a}-\gamma+\nu)e^{\frac{1}{2}t(-\sqrt{a}-\gamma-\nu)}}{2\left(\frac{1}{4}(-\sqrt{a}-\gamma-\nu)^2(b)+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(-\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu\right)} -$$
$$\frac{\eta\nu(-\sqrt{a}-\gamma+\nu)e^{\frac{1}{2}t(\sqrt{a}-\gamma-\nu)}}{2\left(\frac{1}{4}(\sqrt{a}-\gamma-\nu)^2(3\eta+3\gamma+3\nu)+\frac{1}{2}(\sqrt{a}-\gamma-\nu)(c)-\alpha\eta\gamma+\frac{1}{2}(\sqrt{a}-\gamma-\nu)^3+\eta\gamma\nu\right)} + \frac{\eta\gamma\nu}{\eta\gamma\nu-\alpha\eta\gamma}$$

$$J = 1 - \frac{e^{-\eta t}(-\sqrt{a}+2\eta-\gamma-\nu)(\sqrt{a}+2\eta-\gamma-\nu)}{4\left(-\alpha\gamma+\eta^2-\eta\gamma-\eta\nu+\gamma\nu\right)}$$

$$K = -\frac{\eta(\sqrt{a}-\gamma-\nu)(\sqrt{a}+\gamma+\nu)}{4(\eta\gamma\nu-\alpha\eta\gamma)}$$

$$e^V = \begin{pmatrix} A & B & 0 & 0 \\ C & D & 0 & 0 \\ E & F & G & 0 \\ H & I & J & K \end{pmatrix}$$

B.1.6 Choosing Parameters for the Huisman Method when Fitting to Simulated Data

To choose a shedding load profile, we first re-fit a spline to the points from the Nourbakhsh et al. [2022] profile raised to the tenth power, with an additional value of 0 at time 0. We then generated predictions from the spline from 0 to 29 evenly spaced by 0.1, and used these as true values from the shedding load profile. We then used the Nelder-Mead algorithm to search for shape and scale parameters of the gamma distribution that minimized the squared loss of the proposed grid point values versus our generated true values, and used these parameters as the shape and scale of the shedding load profile for the Huisman model. In an SEIR model, the intrinsic generation time distribution is the sum of the latent and infectious periods [Svensson, 2007, Champredon and Dushoff, 2015, Champredon et al., 2018], which is a hypo-exponential distribution. `EpiEstim` is normally used assuming the generation time distribution is a gamma distribution. We used a gamma distribution with mean and standard deviation equal to the true intrinsic generation time distribution.

B.1.7 Branching Process Inspired Models

For additional comparisons of our estimates of SARS-CoV-2 in Los Angeles, CA, we use two branching process inspired models that, unlike compartmental models, only model latent incidence. The Huisman et al. [2022b] method uses one example of this class of methods, but there are many others. The method relies on the so-called renewal equation that calculates current incidence as a product of a weighted sum of previous incidence and the effective reproduction number. Let I_t be the incidence at time t, R_t be the effective reproduction number at time t, and $g(t)$ be the probability density function of the generation time distribution (the time between an individual becoming infected and infecting another individual; under the compartmental model framework this is usually taken to be equivalent to the sum of the latent period and the infectious period [Svensson, 2007, Champredon and Dushoff, 2015, Champredon et al., 2018]). Then the classic renewal equation is:

$$E[I_t|I_1, \dots, I_{t-1}] = R_t \sum_{s=1}^{t-1} I_s g(t-s).$$

The `epidemia` package can be used to create different branching process inspired models to estimate the effective reproduction number using different observation models and models for latent incidence [Scott et al., 2021]. For the model we used in this chapter, we modeled observed cases using a negative binomial distribution, modeled the effective reproduction number as a Gaussian random walk, and modeled unobserved incidence as an auto-regressive normal random variable with variance equal to the mean multiplied by an over-dispersion parameter. The explicit model is listed below:

$$
\begin{aligned}
\tau &\sim \exp(\lambda)\text{–Hyperprior for unobserved incidence,}\\
I_\nu &\sim \exp(\tau)\text{–Prior on unobserved incidence } \nu \text{ days before observation,}\\
I_{\nu+1}, \dots, I_0 &= I_\nu\text{–Unobserved incidence,}\\
\sigma &\sim \text{Truncated-Normal}(0, 0.1^2)\text{–Prior on variance of random walk}\\
\log R_0 &\sim \text{Normal}(\log 2, 0.2^2)\text{–Prior on } R_0\text{,}\\
\log R_t | \log R_{t-1} &\sim \text{Normal}(\log R_{t-1}, \sigma)\text{–Random walk prior on } R_t\text{,}\\
\psi &\sim \text{Normal}(10, 2)\text{–Prior on variance parameter for incidence,}\\
I_t | I_\nu, \dots, I_{t-1} &\sim \text{Normal}(R_t \sum_{s<t} I_s g_{t-s}, \psi)\text{–Model for incidence,}\\
\alpha &\sim \text{Normal}(0.13, 0.7^2)\text{–Prior on case detection rate,}\\
y_t &= \alpha_t \sum_{s<t} I_s \pi_{t-s}\text{–Mean of observed data model,}\\
\phi &\sim P(\phi)\text{–Prior on dispersion parameter for observed data,}\\
Y_t &\sim \text{Neg-Binom}(y_t, \phi)\text{–Observed data model.}
\end{aligned}
$$

Here π_t are the values of the probability density function for the delay distribution, the time between an individual being infected and being observed. We also used the Rt-estim-gamma model, which is similar to the model above, but uses total diagnostic tests as a model covariate in the observation model. This allows the rate of detection to change over time as a function of available tests, avoiding the situation, for example, where an increase in cases due to test availability is mistaken for an increase in cases due to increased incidence. Full details are available in [Goldstein et al., 2024].

B.2 Results

B.2.1 Simulating an Agent-Based Stochastic SEIRR Model

An agent-based stochastic SEIRR model is a continuous time Markov jump process on the cartesian state-space $\{S, E, I, R1, R2\}^N$. When represented as a vector $\mathbf{G}(t)$, each entry of $\mathbf{G}(t)$ records the state of one of the N individuals, i.e. if the ith entry of $\mathbf{G}(t)_i$, $G(t)_i$ is S, then the ith individual is susceptible at time t. Let $I(t)$ be the number of infectious individuals at time t. The Markov jump process can be defined in terms of its transition rates from state $\mathbf{G}$ to state $\mathbf{G}'$ so that

$$\lambda_{\mathbf{G}\mathbf{G}'} = \begin{cases} \beta_t/N \times I(t) \text{ if } G_j = S \text{ and } G'_j = E, \\ \gamma \text{ if } G_j = E \text{ and } G'_j = I, \\ \nu \text{ if } G_j = I \text{ and } G'_j = R1, \\ \eta \text{ if } G_j = R1 \text{ and } G'_j = R2, \\ 0 \text{ otherwise.} \end{cases}$$

The well known Gillespie algorithm popularized in [Gillespie, 1977] can be used to simulate from this model, but it is quite slow. We employ a variation of the Gillespie algorithm in order to simulate individuals and their individual state transition times. In essence, as an individual enters the simulation (via infection) all future transition times for that individual are simulated at once. Then, the next event is simply the most recent transition time amongst all individuals still in the simulation. In psuedocode, the basic algorithm is

Algorithm 1 Individual Gillespie Algorithm

```
$i \leftarrow$ initial I
$e \leftarrow$ initial E
$r1 \leftarrow 0$
$s \leftarrow N - i - e$
$t_i \leftarrow$ Initial infectious times
$t_{r1} \leftarrow$ Initial recover times
$t_{r2} \leftarrow$ Initial stop shedding times
$t \leftarrow \max t_i$
while $i > 0 | r1 > 0$ do
    if $s > 0 \& i > 0$ then
        $n_e \leftarrow t + \text{Exp}(\beta \times s \times i)$
    else if $s = 0 | i = 0$ then
        $n_e \leftarrow \inf$
    end if
    $n_i \leftarrow \min t_i$
    $n_{r1} \leftarrow \min t_{r1}$
    $n_{r2} \leftarrow \min t_{r2}$
    $n_t \leftarrow \min n_e, n_i, n_{r1}, n_{r2}$
    $t \leftarrow n_t$
    if $n_t = n_e$ then
        $s \leftarrow s - 1$
        $e \leftarrow e + 1$
        $x \sim \text{Exp}(\gamma)$
        $y \sim \text{Exp}(\nu)$
        $z \sim \text{Exp}(\eta)$
        $t_i \leftarrow \text{append}(t_i, t + x)$
        $t_{r1} \leftarrow \text{append}(t_{r1}, t + x + y)$
        $t_{r2} \leftarrow \text{append}(t_{r2}, t + x + y + z)$
    else if $n_t = n_i$ then
        $e \leftarrow e - 1$
        $i \leftarrow i + 1$
        $t_i \leftarrow \text{remove}(t_i, n_i)$
    else if $n_t = n_{r1}$ then
        $i \leftarrow i - 1$
        $r1 \leftarrow r1 + 1$
        $t_{r1} \leftarrow \text{remove}(t_{r1}, n_{r1})$
    else if $n_t = n_{r2}$ then
        $r1 \leftarrow r1 - 1$
        $t_{r2} \leftarrow \text{remove}(t_{r2}, n_{r2})$
    end if
end while
```

We omit the various pieces of code that ensure that all of the individual transition times are recorded and associated with the correct individuals. It is easy to adapt this algorithm to allow for a changing β at known change point times. We simply add an additional check, where the next event can be any of the nearest future transition times, or the nearest future change point time. We use this adapted version when we simulate data to test the EIRR-ww model. In order to calculate individual genome concentrations, we use the consensus shedding profile for SARS-CoV-2 RNA created by Nourbakhsh et al. [2022]. We manually recorded the values displayed in the figure, and then used a thin plate regression spline (using the `R` package `fields` [Douglas Nychka et al., 2021]) to create a continuous curve. The spline fit and manually recorded values are displayed in Figure B.1.

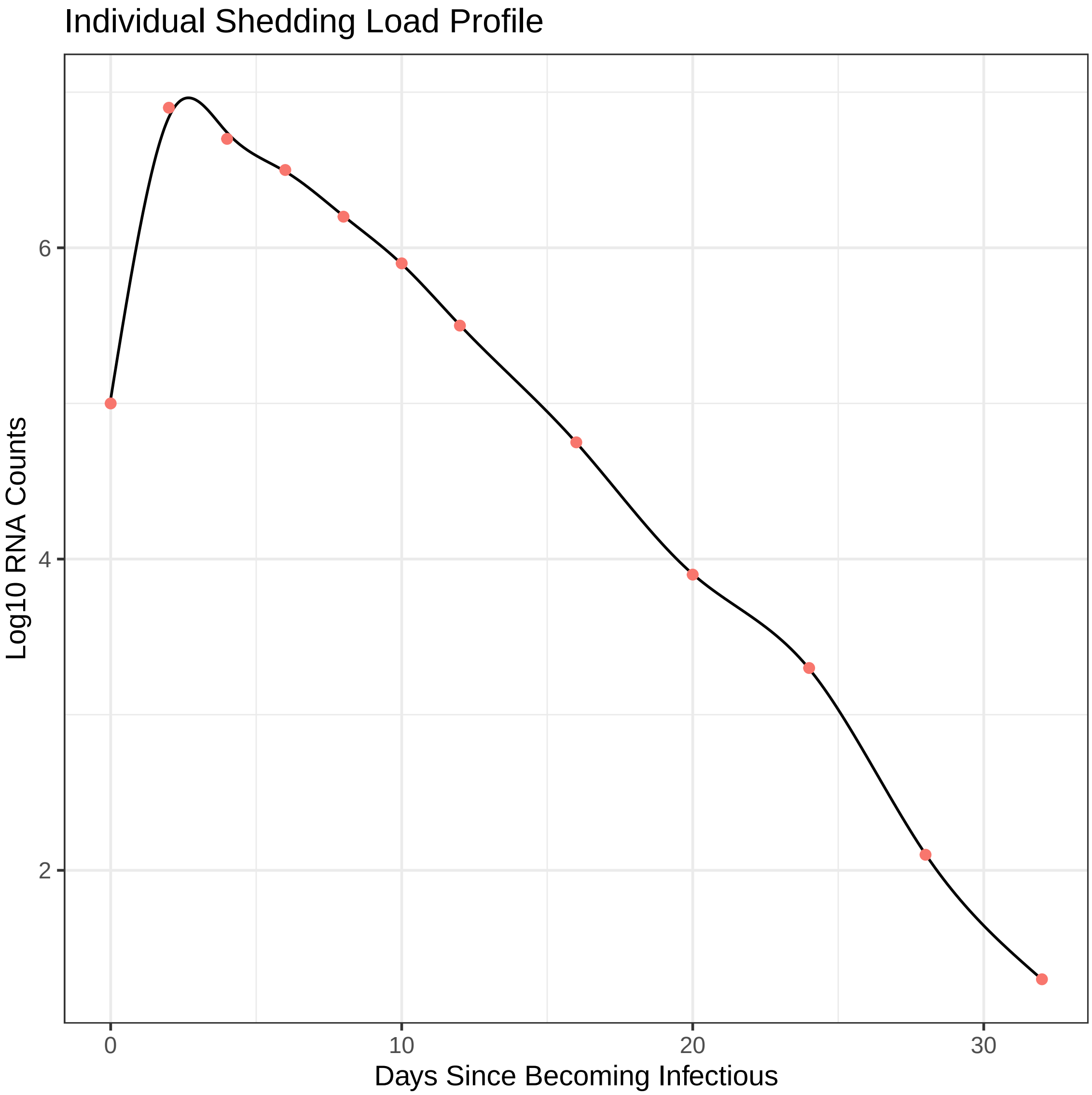


Figure B.1: Thin plate spline fit to manually recorded values from the consensus shedding load profile developed by Nourbakhsh et al. (2022). Red dots are the manually recorded values. Black line is the spline fit.

To calculate the mean log genome concentration shed by an individual at time time l, we first checked to see if they were in the I or $R1$ compartments at time l, then predicted mean log genome concentration using the fitted spline and the difference between l and the time they became infectious.

B.2.2 Comparing Simulation Engines

We compare our variation on the Gillespie algorithm to both a traditional Agent based model (Agent) and a compartmental model (Compartment) simulated using the traditional Gillespie algorithm. We set the population to be 100, set R_0 to 1.5 and initialized with 5 infectious individuals and all other individuals in the susceptible population. We simulated 10000 simulations from each of the three engines, and plotted the mean and quantiles of the counts in each of the compartments in the first 100 days. For the agent and variation models, we also plotted the log concentrations on each of the first 100 days. We see no evidence the three engines are not equivalent.

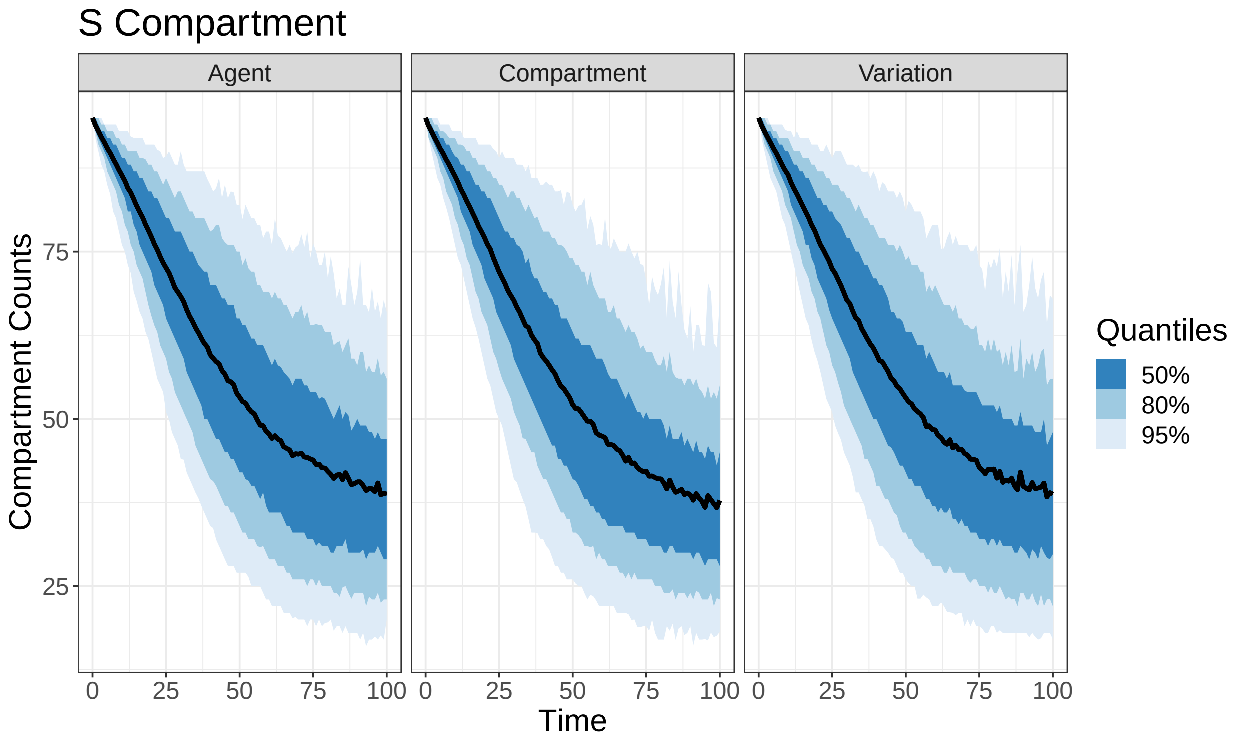


Figure B.2: Counts in the S compartment across three simulation engines. Blue bars are quantiles, black lines are means.

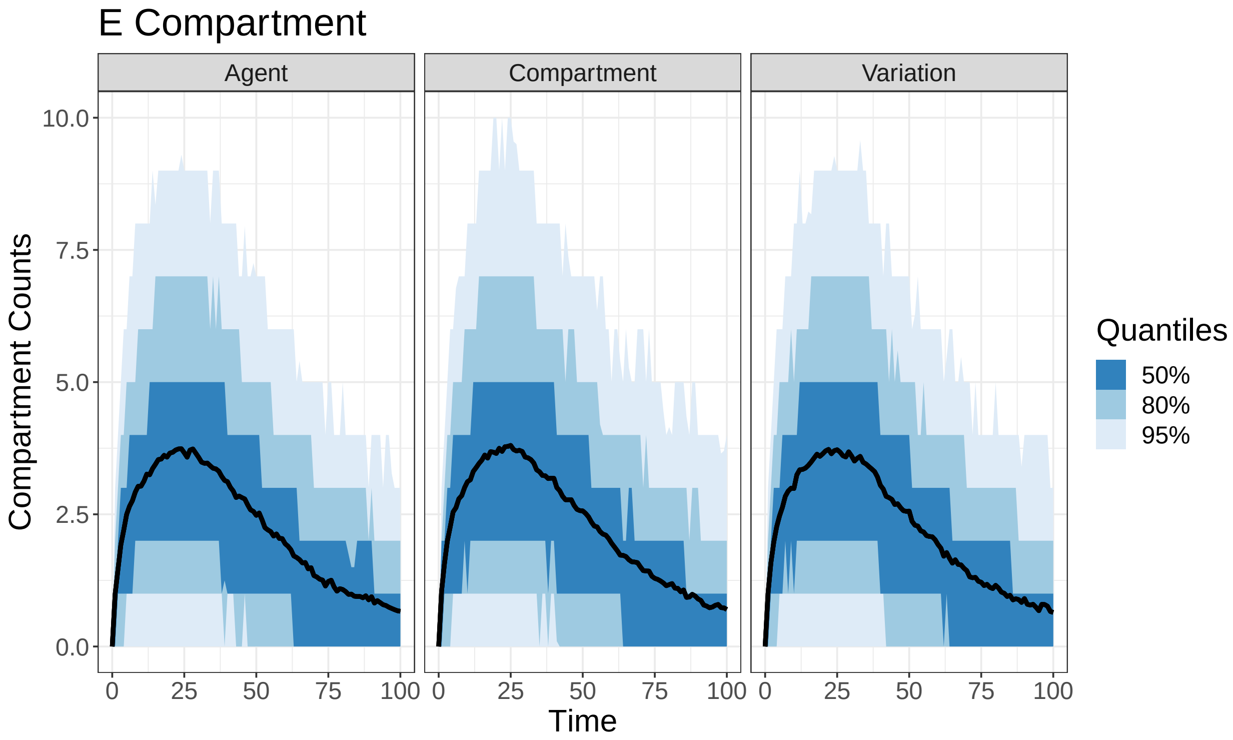


Figure B.3: Counts in the E compartment across three simulation engines. Blue bars are quantiles, black lines are means.

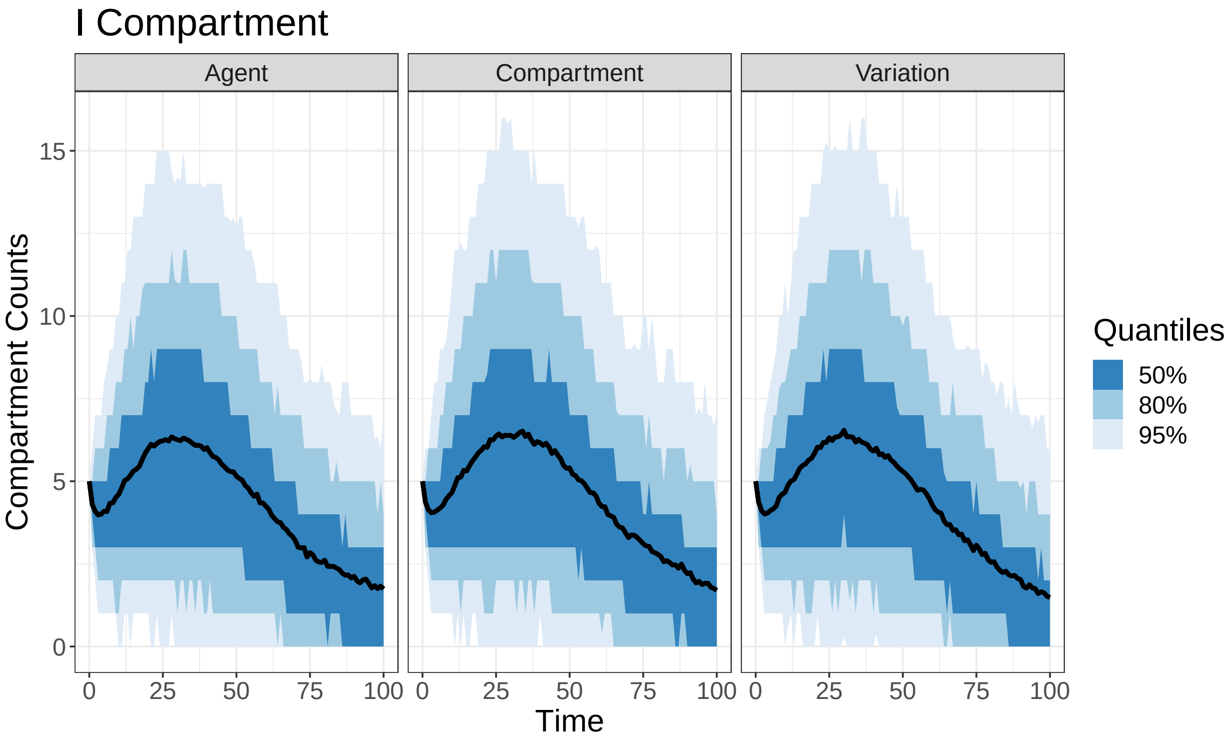


Figure B.4: Counts in the I compartment across three simulation engines. Blue bars are quantiles, black lines are means.

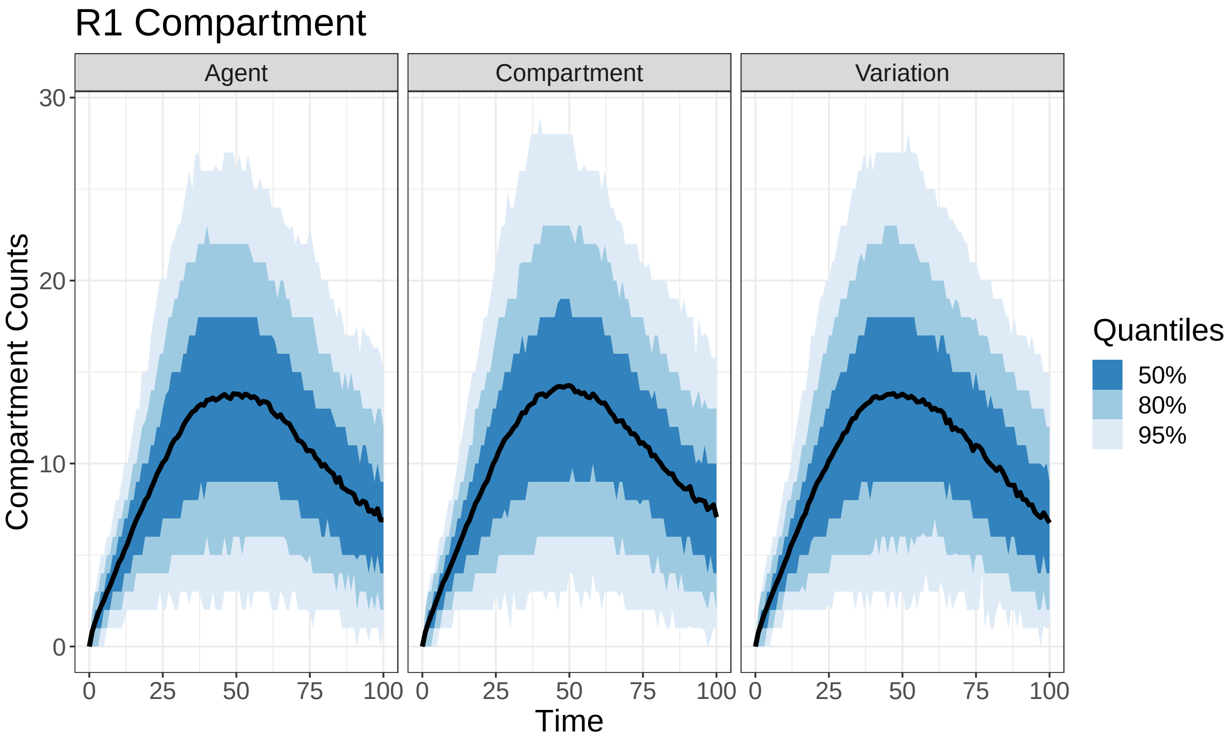


Figure B.5: Counts in the R1 compartment across three simulation engines. Blue bars are quantiles, black lines are means.

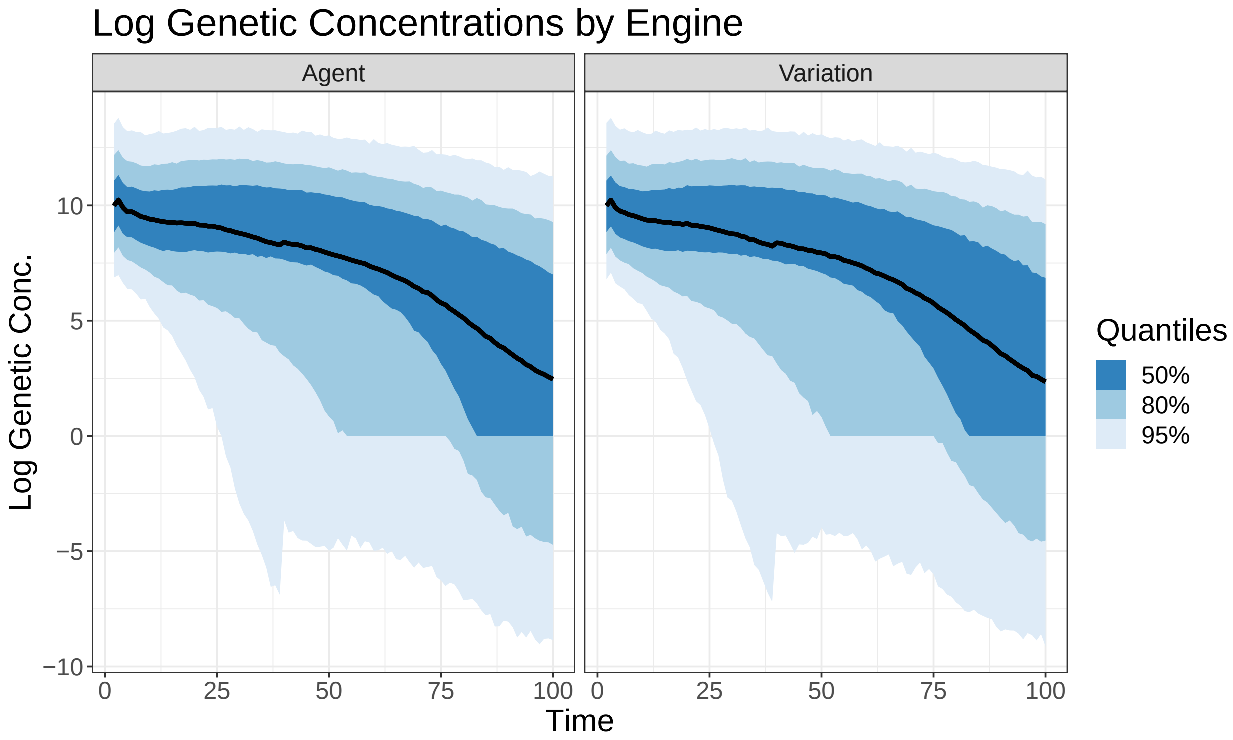


Figure B.6: Simulated log concentrations across two simulation engines. Blue bars are quantiles, black lines are means.

B.2.3 Simulation Parameters

For choosing the prior for λ, the population was 1000, R_0 was set to 2, and the parameters for γ, ν and η were drawn from the priors in Table B.2. Five individuals were initially in the I compartment, the other individuals were in the S compartment.

The parameters used for the simulation producing observed data used in the simulation studies described in the main text are displayed below. The population size was 100000, the model was initialized with 200 individuals in the E compartment and 200 in the I compartment, all others were in the S compartment. The basic reproduction number was set to 1.75 for five weeks, 0.7 for four weeks, and then increased from 0.7 to 0.9 for two weeks before following the trajectory chosen for the observed data period.

Parameter	Interpretation	Value
$1/\gamma$	Mean latent period duration	4
$1/\nu$	Mean infectious period duration	7
$1/\eta$	Mean duration when recovered but still shedding RNA	18
ρ	scales concentrations of RNA into observed concentrations	0.011
τ	describes the noisiness of observed gene data	0.5

Table B.1: Simulation parameters used in the simulation study. Durations are measured in days.

The estimate for $1/\gamma$ was calculated by averaging the mean latent period calculated by Xin et al. [2022] with the mean time to detecting virus that could be cultured found in [Killingley et al., 2022]. We took culturable virus to be a proxy for infectiousness. The mean time from infectiousness to symptom onset was 1.37, calculated again as an average from the previous two studies (1.4 days from [Xin et al., 2022] versus 1.33 from [Killingley et al., 2022]). Mean infectious period was calculated using the mean period of detecting virus that could be cultured found in [Killingley et al., 2022]. Many studies have calculated the time from symptom onset to the end of RNA shedding in fecal matter; we averaged the estimates from [Okita et al., 2022] and [Zhang et al., 2021]. Because these two literature reviews shared studies, we dis-aggregated their estimates into individual study estimates, counting each study only once in our final average. We used mean shedding estimates from each study reported by Okita et al. [2022]. Zhang et al. [2021] did not include estimates of the mean for each study in their review, if estimates of the mean duration were available from the original paper, they were used, if they were not, the paper was not included in the final average. We also examined the literature review by Walsh et al. [2020], but found no new studies with more than two samples not in the previous two reviews. We decided to exclude studies with fewer than three samples. The final average duration from symptom onset to the end of RNA shedding in fecal matter was 22.99 days. We calculated $1/\eta$ (the mean duration of shedding after recovery) as

$$1/\eta = 22.99 + 1.33 - 1/\nu = 17.86.$$

Note we are assuming shedding begins at the start of the infectious period. Note that all of the parameters are based on studies of the original Wuhan lineage of SARS-CoV-2. To calculate a plausible τ (standard deviation from true genetic concentration), we used the JWPCP data, and fit a Bayesian thin plate regression generalized student t-distribution spline to the data. We used the mean of the posterior estimate of τ for the τ in our simulation. The value of df was chosen as the mean of the posterior estimate from the thin plate spline model. The value for ρ (scaling factor for the mean of the generalized t-distribution) was chosen arbitrarily so that the mean total genome concentrations produced by the simulation would be roughly similar to the means seen in the JWPCP data. The value for ϕ (the negative binomial over-dispersion parameter) was chosen by fitting a negative-binomial spline to Los Angeles case data, and using the mean from the posterior estimate for ϕ. The value for ψ (mean case detection rate) we chose based on our study of the spread of SARS-CoV-2 in Orange County, CA, the neighboring county to Los Angeles. [Bayer et al., 2024]. In that study, we estimated a weekly case detection rate, and at the end of the study period we estimated between 1 in 5 and 1 in 7 new infections were being detected. We chose to use 0.2 (1 in 5 cases being detected) for our simulation study.

Table B.2: Priors used by all models in the baseline simulation scenario.

Parameter	Model	Prior	Prior Median (95% Interval)	Truth
γ	All	Log-normal(log(1/4), 0.2)	0.25 (0.17, 0.37)	0.25
ν	All	Log-normal(log(1/7), 0.2)	0.14 (0.10, 0.21)	0.14
σ_{rw}	All	Log-normal(log(0.1), 0.2)	0.1 (0.07, 0.15)	NA
η	SEIRR-ww/EIRR-ww	Log-normal(log(1/18), 0.2)	0.06 (0.04 0.08)	0.06
$R_{0,0}$	SEIRR-ww/SEIR-cases	Log-Normal(log(0.88), 0.1)	0.88 (0.72, 1.07)	0.9
λ	SEIRR-ww/EIRR-ww	Logit-normal(5.69, 2.18)	0.997 (0.81, 1)	NA
τ	SEIRR-ww/EIRR-ww	Log-normal(0, 1)	1.00 (0.14, 7.10)	0.5
df	SEIRR-ww/EIRR-ww	Gamma(10,2)	19.33 (9.59, 34.17)	2.99
ρ	SEIRR-ww/EIRR-ww	Log-normal(0, 1)	1.00 (0.14, 7.10)	NA
R_0	EIRR-ww/EIR-cases	Log-Normal(log(0.88, 0.1)	0.88 (0.72, 1.07)	0.80
ψ	SEIR-cases/EIR-cases	Logit-Normal(-1.39, 0.4)	0.20 (0.10, 0.35)	0.2
ϕ	SEIR-cases/EIR-cases	Log-Normal(4.22, 0.29)	68.03 (38.54, 120.11)	57.55
S_SEIR1	SEIRR-ww	Logit-Normal(3.47, 0.05)	0.97 (0.967, 0.972)	0.97
I_EIR1	SEIRR-ww	Logit-Normal(-1.548302, 0.05)	0.18 (0.16, 0.19)	0.18
$R1\_ER1$	SEIRR-ww	Logit-Normal(2.22, 0.05)	0.90 (0.89, 0.91)	0.90
S_EI	SEIR-cases	Logit-Normal(4.83, 0.05)	0.992 (0.991, 0.993)	0.992
I_EI	SEIR-cases	Logit-Normal(0.78, 0.05)	0.68 (0.66, 0.71)	0.68
$E(0)$	EIRR-ww/EIR-cases	Normal(225, 0.05)	225.00 (224.90, 225.01)	225
$I(0)$	EIRR-ww/EIR-cases	Normal(489, 0.05)	489.00 (448.90, 489.01)	489
$R1(0)$	EIRR-ww	Normal(2075, 0.05)	2075.00 (2074.90, 2075.01)	2075

Figure B.7 displays the prior quantiles of the random walk prior on R_t for the baseline simulation scenario for the EIRR-ww model.

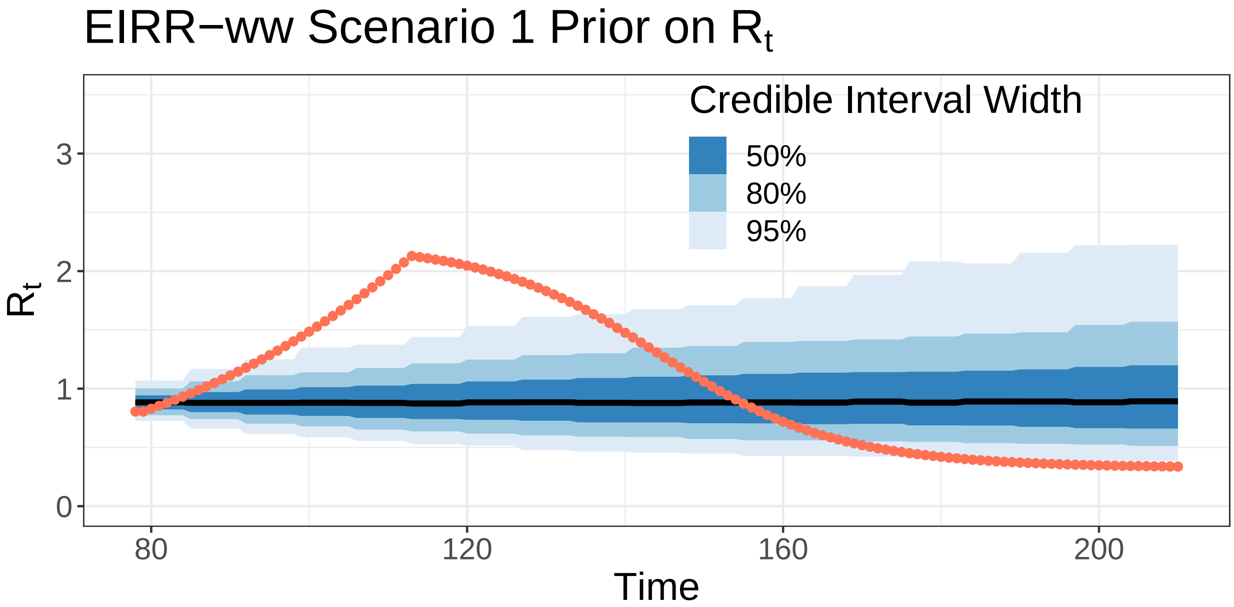


Figure B.7: EIRR-ww prior summaries for the time-varying effective reproduction number. Blue bars are prior credible intervals, black lines are medians, true values are shown in red.

B.2.4 Realization of the Simulation

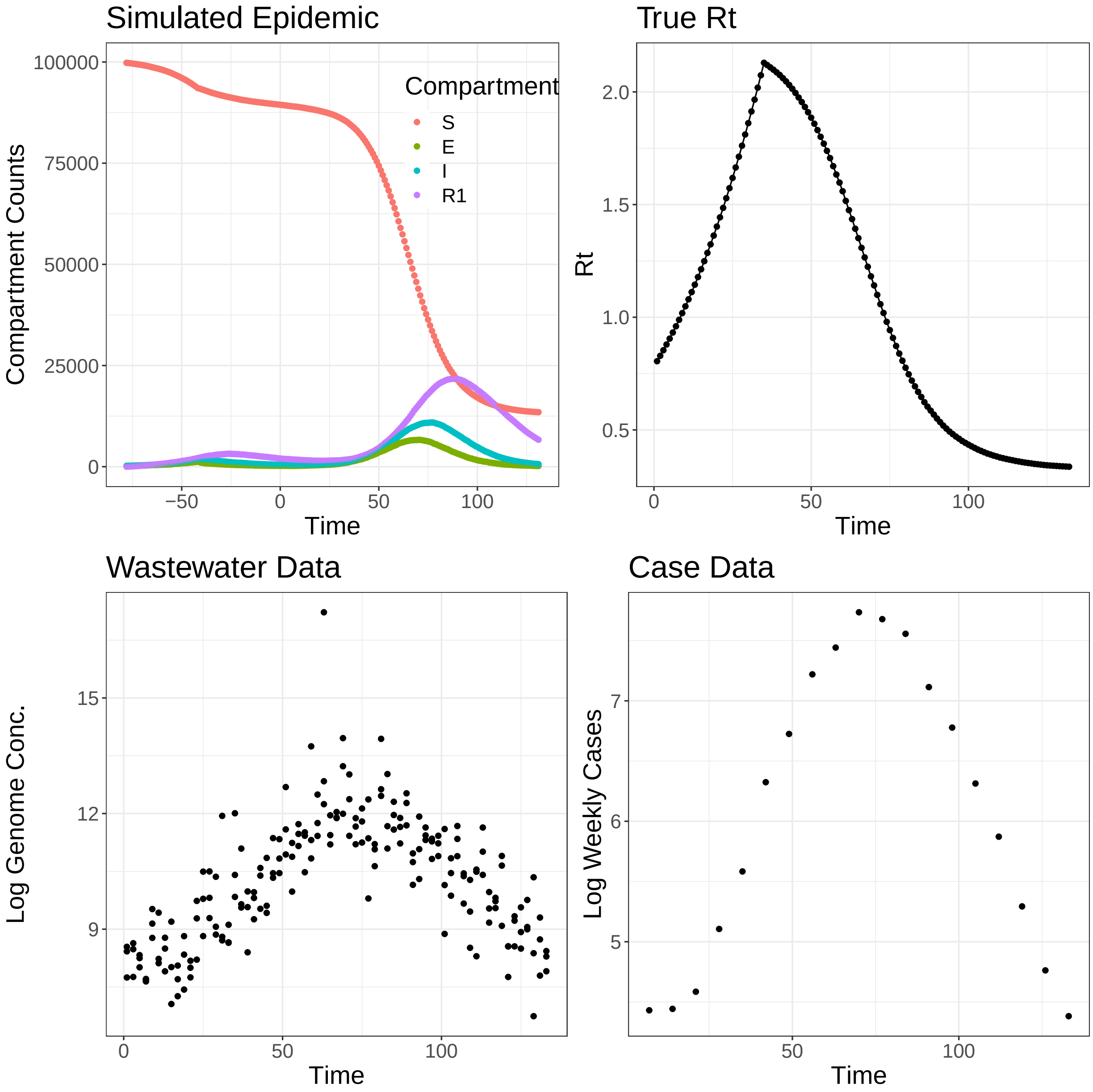


Figure B.8: An example of simulated epidemic data generated from a single realization of a stochastic SEIRR model. Top left panel shows the simulated epidemic, time 0 is the day before the first wastewater sample/case is observed. Top right panel shows the true effective reproduction number trajectory. Bottom row shows simulated wastewater and case data on the log scale.

B.2.5 Computational Considerations

The systems of ordinary differential equations for the EIR/EIRR models are linear, and thus can be solved in closed form. We conducted informal tests on a Macbook Air M2 (2022) to compare performance of using the EIRR-ww model with the closed form solution we derived vs the Tsit5 solver [Tsitouras, 2011] implemented in the `Julia` package `DifferentialEquations` [Rackauckas and Nie, 2017]. First, we compared the EIRR-ww model fit to one realization of the baseline simulation scenario using the closed form solution versus the EIRR-ww model using the ODE solver. Next, we created a new data set using the same dynamics as the baseline scenario, but where wastewater data was sampled every twelve hours, so that the solution to the ODE system had to be solved every half-day as opposed to every day. The results of these experiments are displayed in Table B.3.

Setting	Closed Form Time	Numerical Solver Time
Fitting to day data	855 seconds	872 seconds
Fitting to half-day data	3567 seconds	4260 seconds

Table B.3: Comparing performance of closed form solution to ODE system versus an ODE solver.

Although the computational benefit of using the closed form ODE solution is not substantial, not having to solve ODEs numerically has other advantages. Using an ODE solver inside an MCMC algorithm is a non-trivial undertaking, as tuning parameters controlling the tolerances for solver errors must be specified by the users [Timonen et al., 2022]. Using the closed form solution, our informal testing suggests our computation times are at least as good as the ODE solver, and regardless of computation time, our method is more robust as it does not require properly specifying error tolerance tuning parameters.

B.2.6 EIRR-ww model Performance in Scenario 1

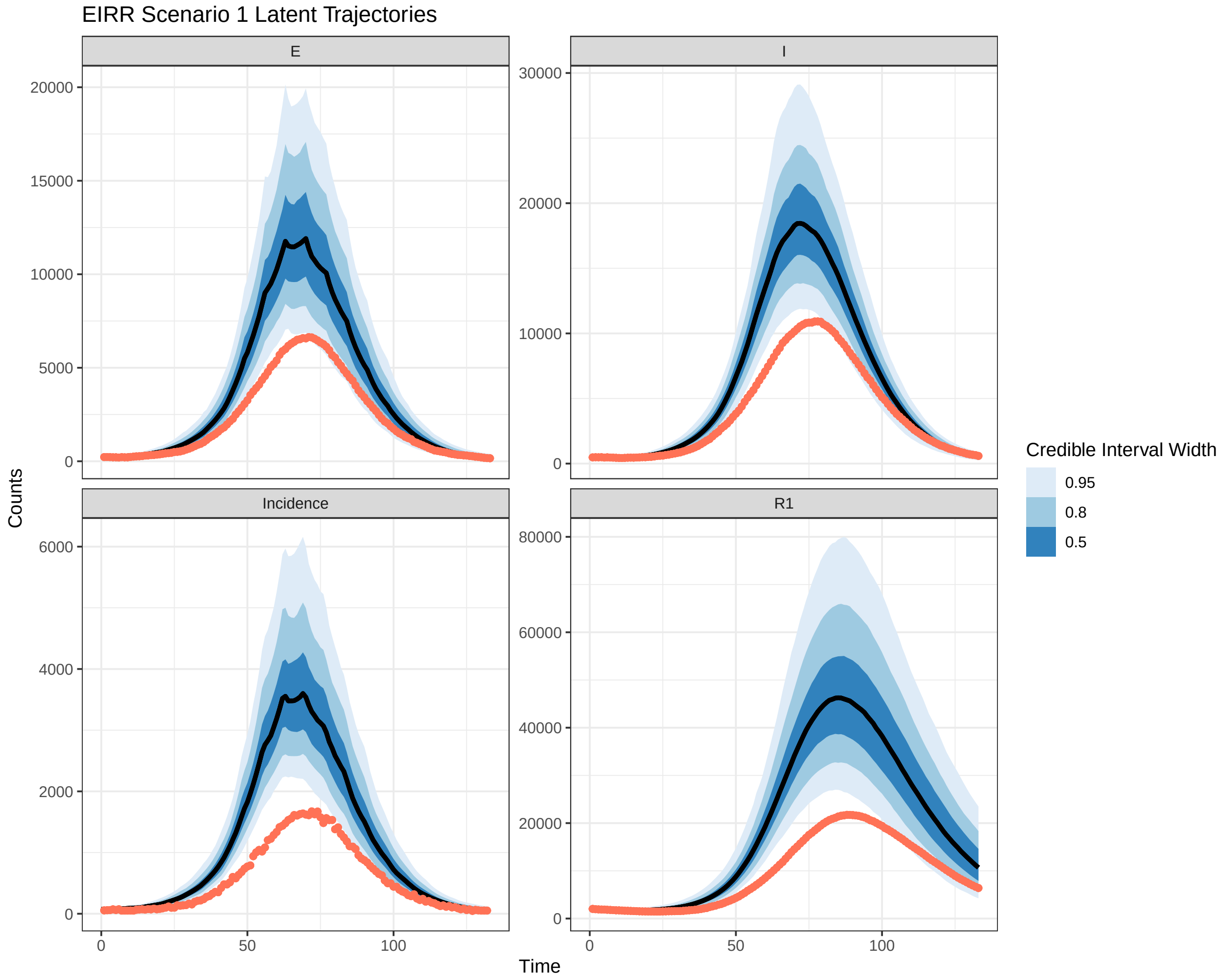


Figure B.9: EIRR-ww posterior summaries for latent unobserved compartments and incidence. Posterior is from the model fit shown in Figure 2. Blue bars are credible intervals, black lines are medians, true values are shown in red.

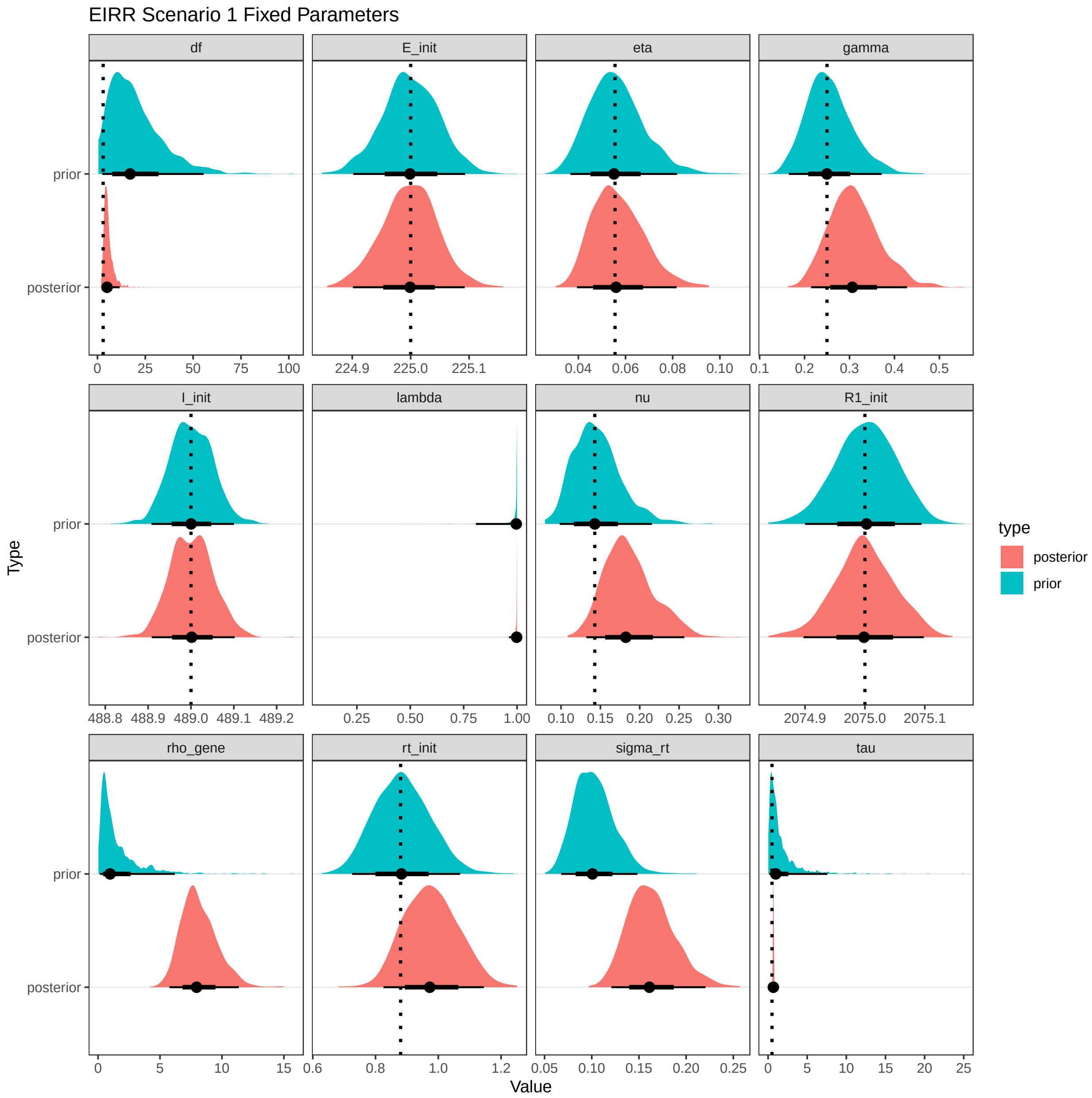


Figure B.10: EIRR-ww prior and posterior summaries for fixed model parameters. Posterior summaries are from the model fit shown in Figure 2. Blue densities are the prior, red densities are the posterior, dotted lines indicate true values (when relevant).

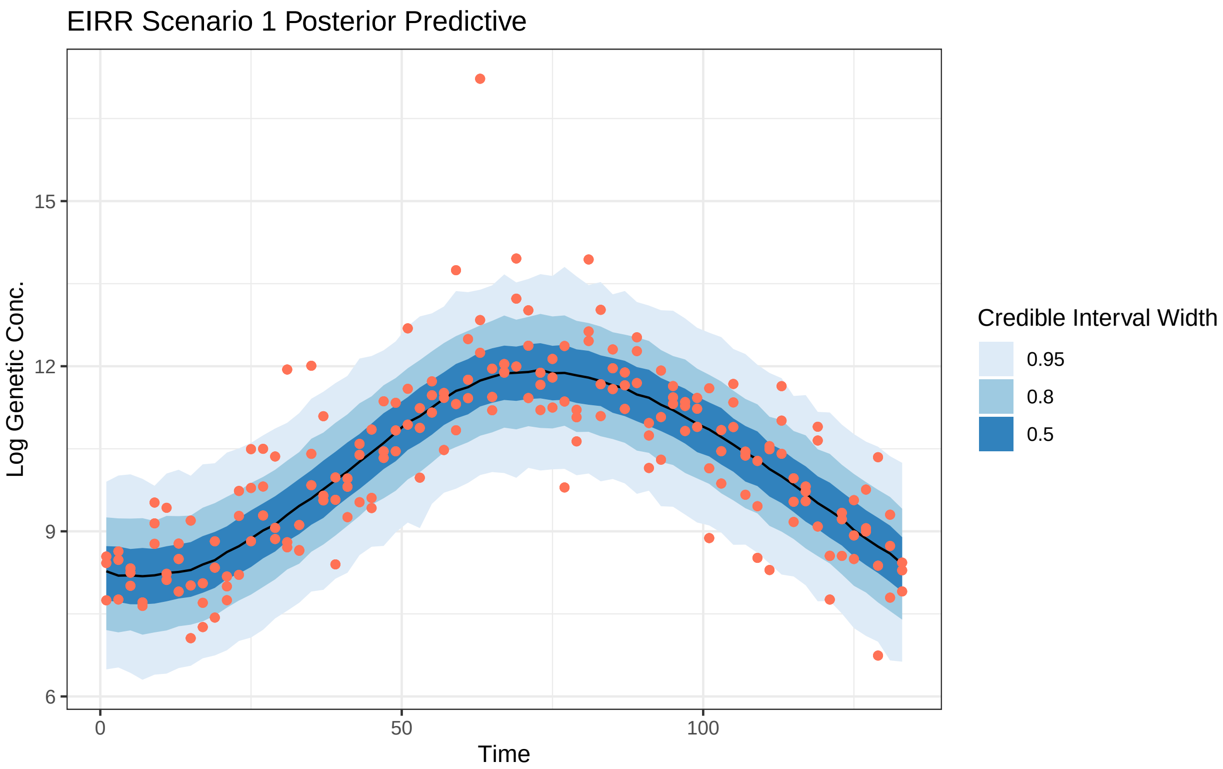


Figure B.11: EIRR-ww posterior predictive of Log genome concentrations. Posterior predictive summaries are from the model fit shown in Figure 2. Blue bars represent credible intervals, black lines medians, and red dots are observed data.

B.2.7 Model Performance Using Finer Random Walk Grid

We used a grid size of 7 days in our analyses, but finer grids are possible. Figure B.12 shows a recreation of Figure 1 in the main text using a finer grid size, the end result is smoother posterior estimates, and while there are some differences in inference between the two model fits, they are largely the same. We chose a 7 day grid as it performs well in simulation, reduces computational burdens, and is more resistant to spurious changes caused by high variation in wastewater data.

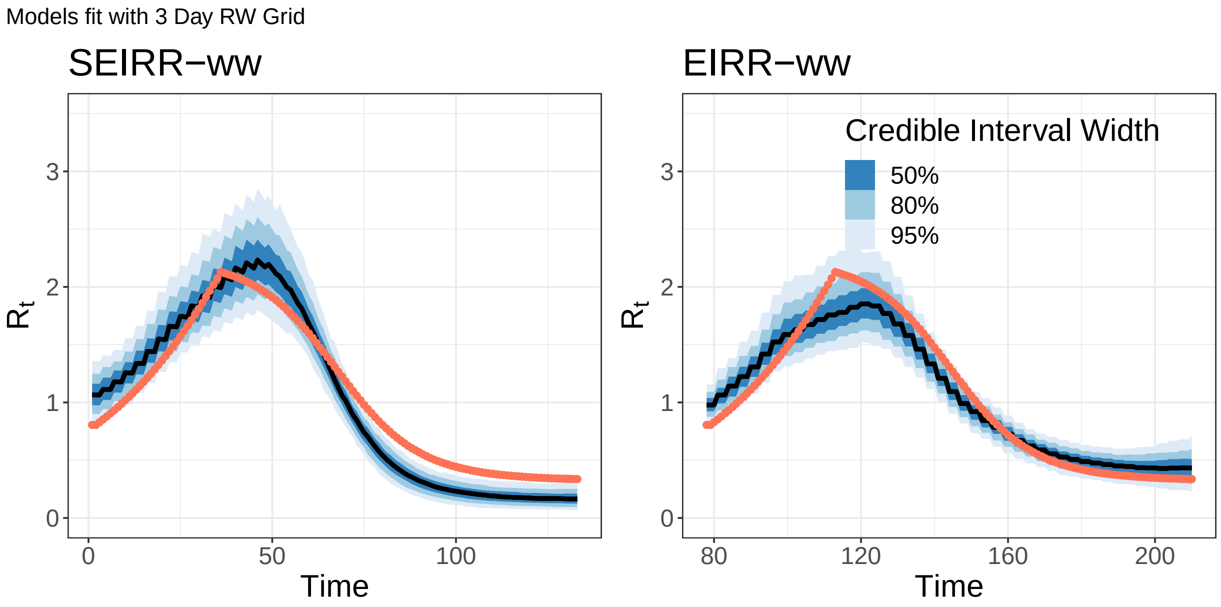


Figure B.12: Posterior summaries of Rt using two models fit to wastewater data using a random walk grid size of 3 days. This is a recreation of Figure 1 in the main text using a grid size of 3 days for the random walk prior, as opposed to 7 days.

B.2.8 MCMC Diagnostics

We used the `posterior` package to calculate $\hat{R}$, ESS bulk and ESS Tail for all parameters for our models in order to assess convergence of MCMC chains Vehtari et al. [2021]. We display the minimum and maximum across all simulations for each of our models and scenarios. We concluded the model had converged if $\hat{R}$ was below 1.05 and the ESS were both above 100. If this was not the case, we re-ran the model for an increased number of iterations (increasing from 500 iterations per chain to 1000 iterations per chain).

Scenario	Max Rhat	Min Rhat	Max ESS Bulk	Min ESS Bulk	Max ESS Tail	Min ESS Tail
EIR-cases	1.03	1.00	7748.92	514.43	4124.70	163.18
EIR-cases LA	1.02	1.00	1213.06	259.83	1060.19	337.05
EIRR-ww (1)	1.03	1.00	4452.76	649.23	2036.34	228.85
EIRR-ww (10 mean)	1.03	1.00	2969.81	300.12	2057.70	114.07
EIRR-ww (10)	1.03	1.00	2176.22	502.38	1133.52	234.43
EIRR-ww (3 mean)	1.03	1.00	5071.60	406.84	2053.79	109.91
EIRR-ww (3)	1.03	1.00	3631.31	528.05	2149.23	226.18
EIRR-ww High Init	1.03	1.00	3342.74	622.39	2042.90	290.15
EIRR-ww LA	1.02	1.00	2249.27	842.62	1129.10	375.09
EIRR-ww Low Init	1.03	1.00	4467.27	425.32	2151.38	126.54
EIRR-ww Low Prop	1.03	1.00	2717.95	436.85	1132.62	100.07
EIRR-ww Stoch Rt	1.03	1.00	3000.00	538.30	1135.72	217.10
SEIR-cases	1.03	1.00	4145.55	414.77	2059.80	167.21
SEIRR-ww	1.03	1.00	4582.79	463.60	2101.94	142.96

B.2.9 Additional Simulation Results

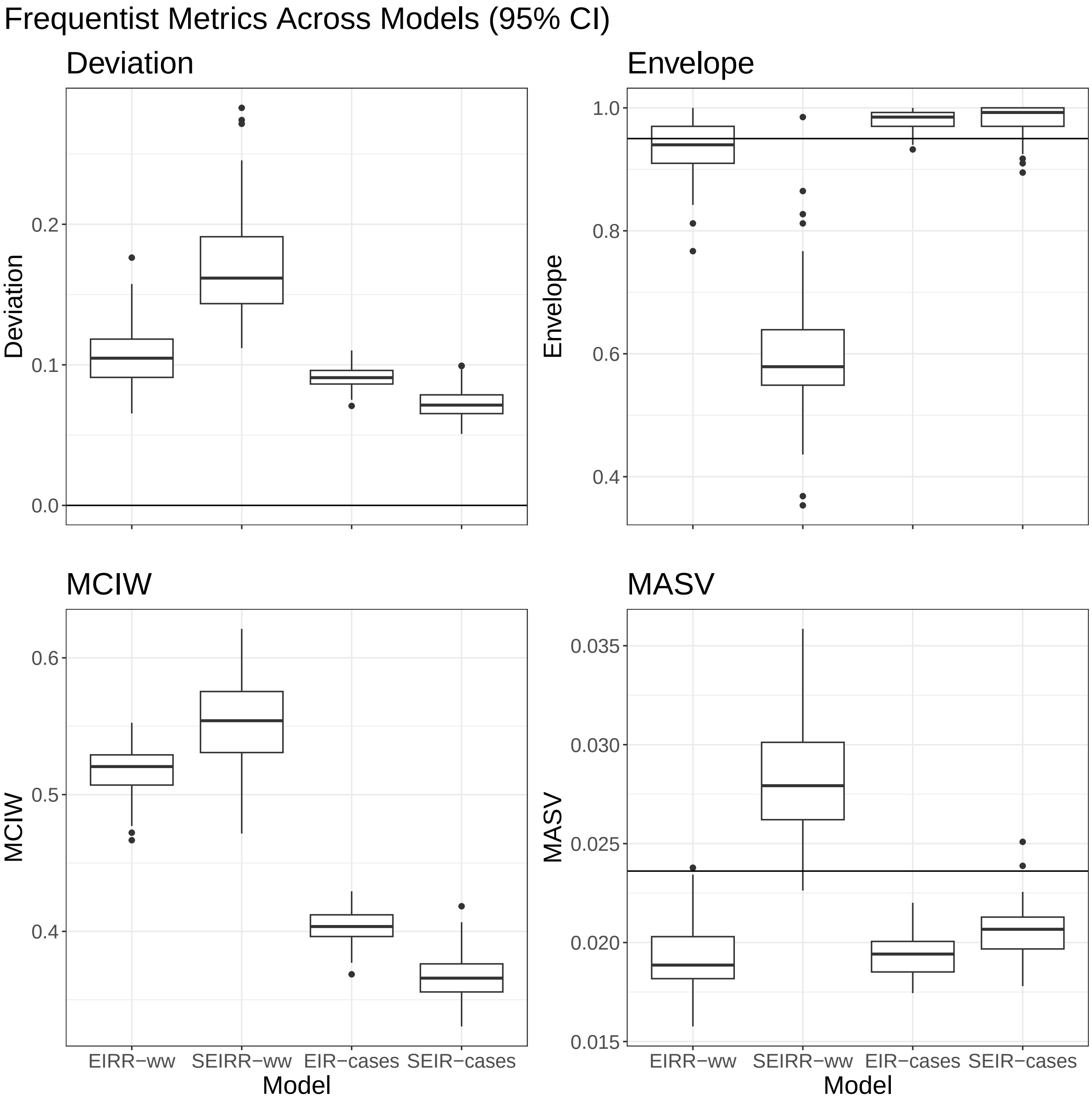


Figure B.13: Frequentist metrics for all models using 95% CI as opposed to 80% CI as in the main text. The ideal Envelope is now 0.95. See Figure 3 in the main text for descriptions of the metrics.

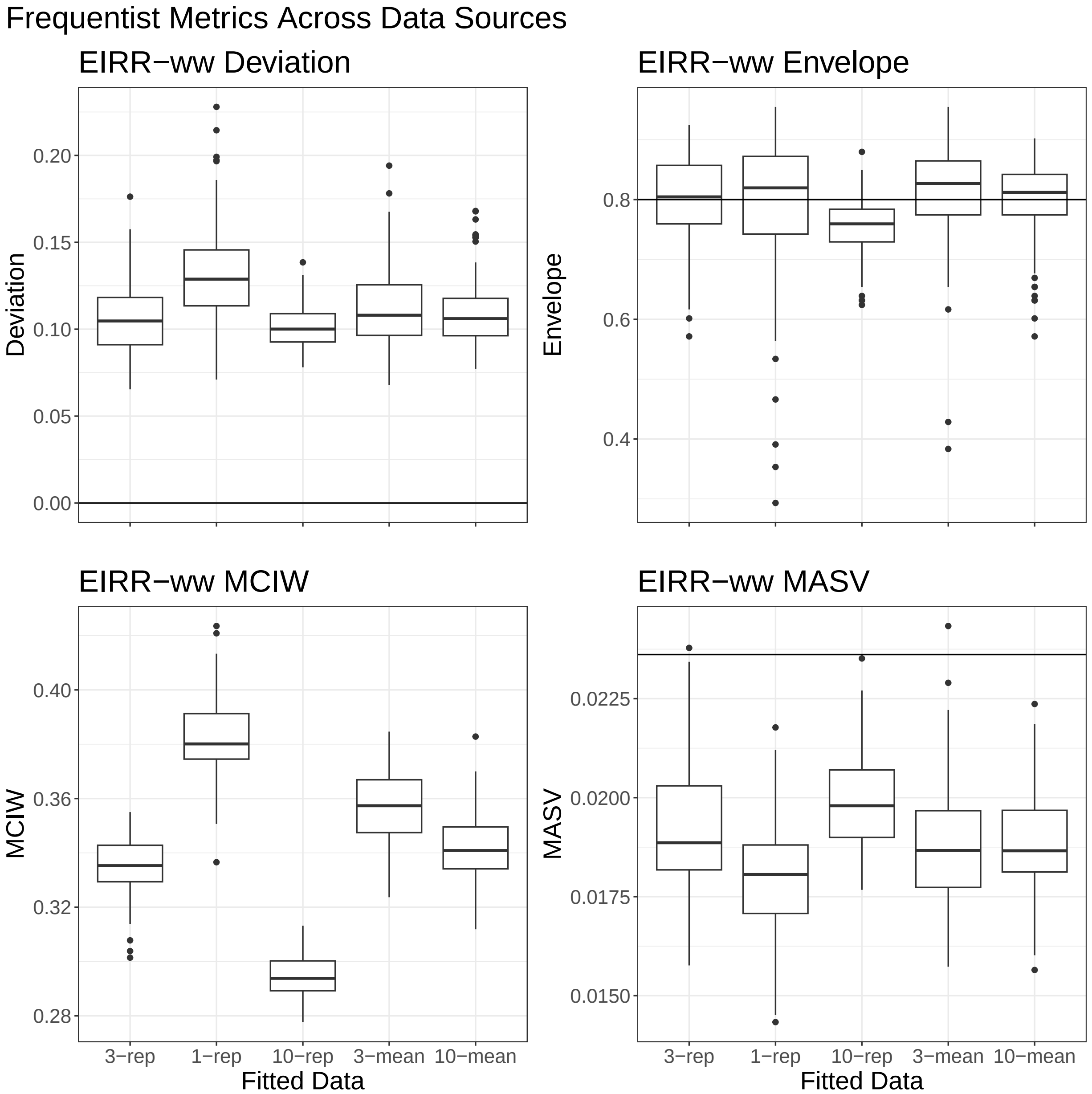


Figure B.14: Frequentist metrics for performance of the EIRR-ww model using different kinds of wastewater data. The baseline scenario 3-reps uses three replicates, 1-rep uses one replicate, 10-reps uses ten replicates, 3-mean uses the mean of three replicates, 10-mean uses the mean of ten replicates. See Figure 3 for descriptions of the metrics.

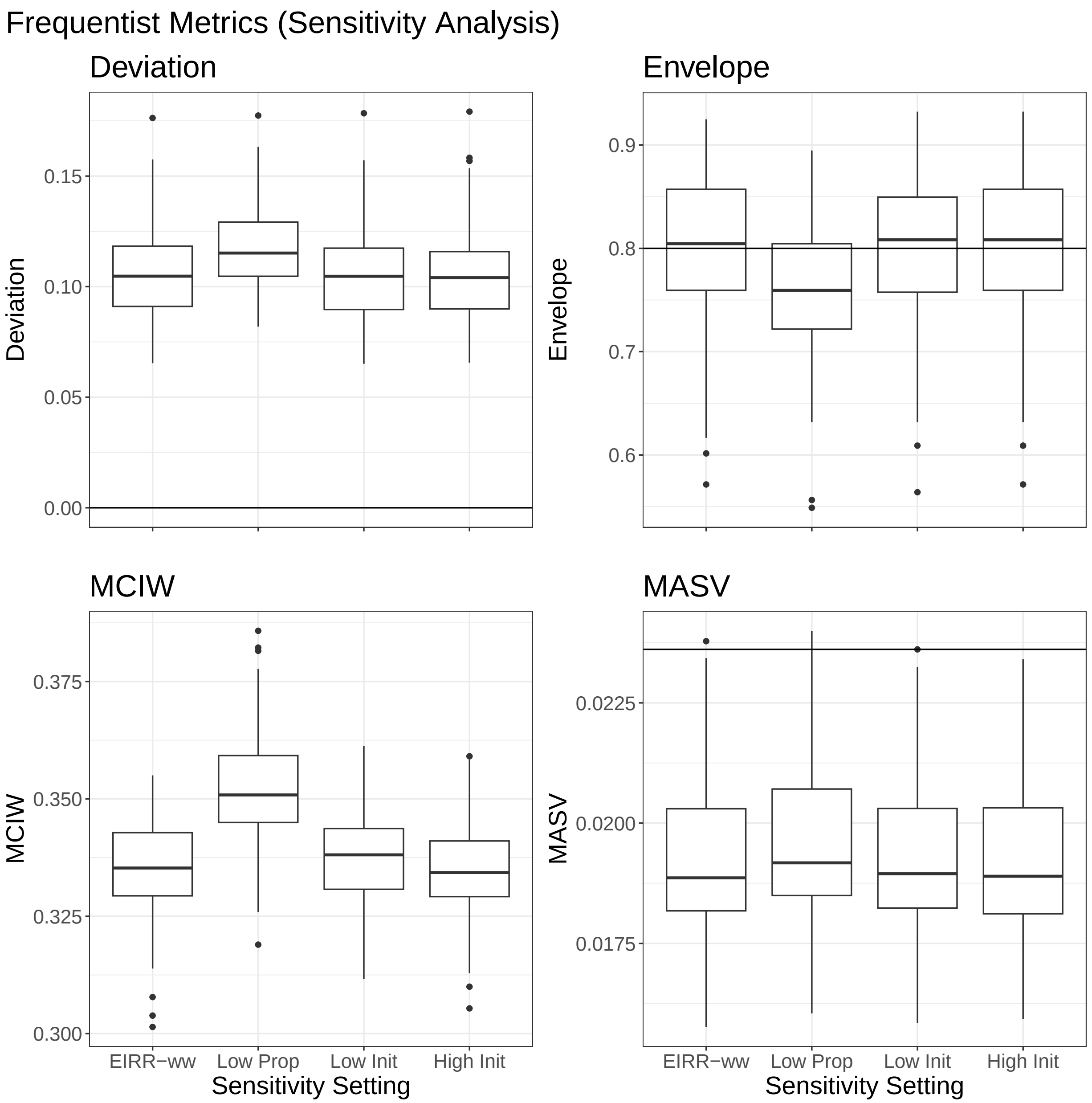


Figure B.15: Frequentist metrics for performance of the EIRR-ww model using mis-specified priors. See Figure 3 in the main text for descriptions of the metrics.

Note that for the Huisman method, because the method relies on generating a synthetic series of cases, the method truncates the inferred R_t values in order to only infer R_t for times for which the method believes there are enough pathogen genome concentrations available to correctly generate synthetic cases. Also, `EpiEstim` does not estimate the effective reproduction number for early points in the time-series. When comparing the Huisman method to the baseline EIRR-ww model, we restrict the comparison to only be on time points for which both models produced inference.

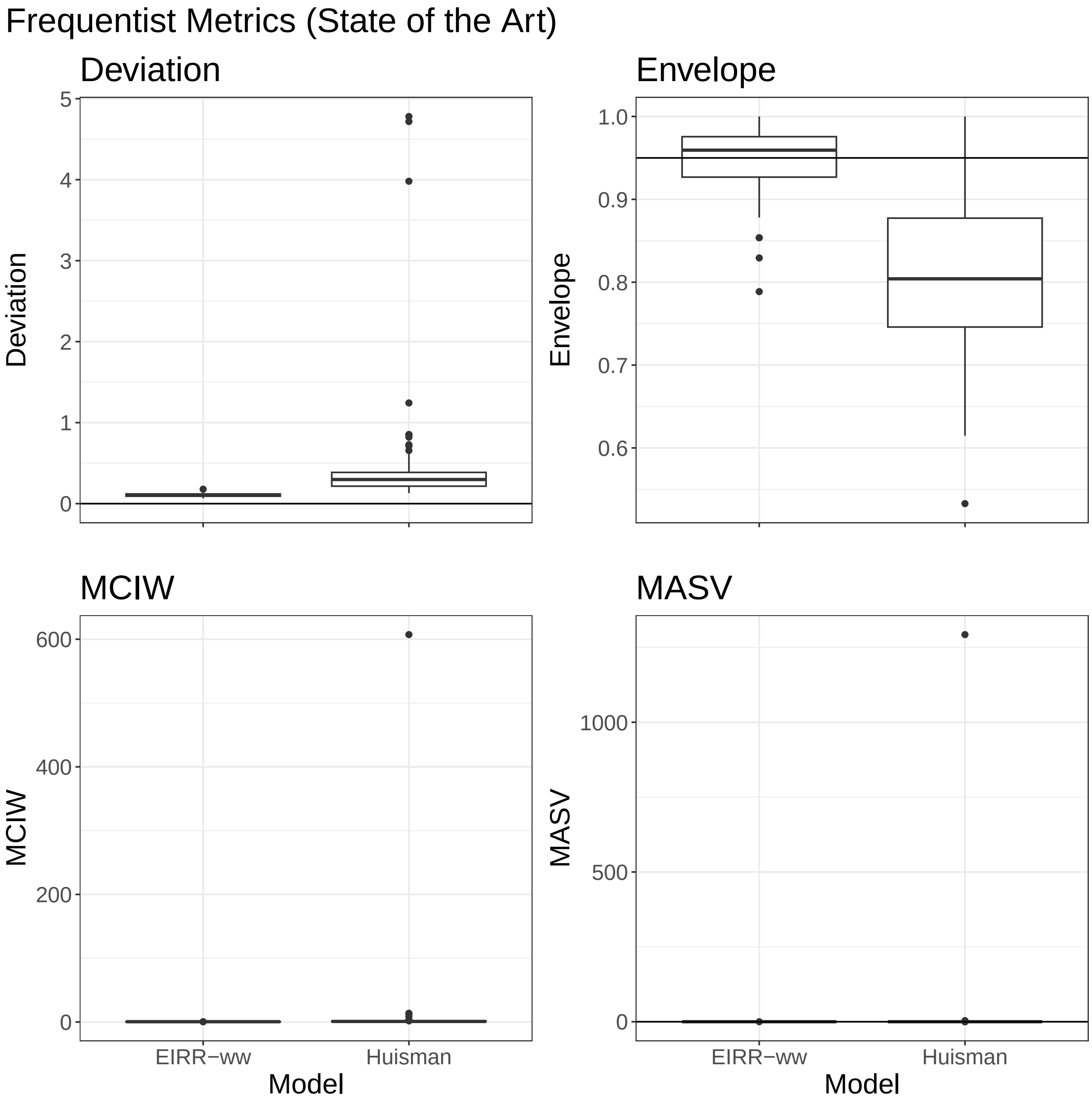


Figure B.16: Frequentist metrics for performance of the EIRR-ww model compared to the Huisman et al. (2022) method. See Figure 3 in the main text for descriptions of the metrics. For one simulation, the deviation of the Huisman model was 600, this was removed from the graph to facilitate visualization.

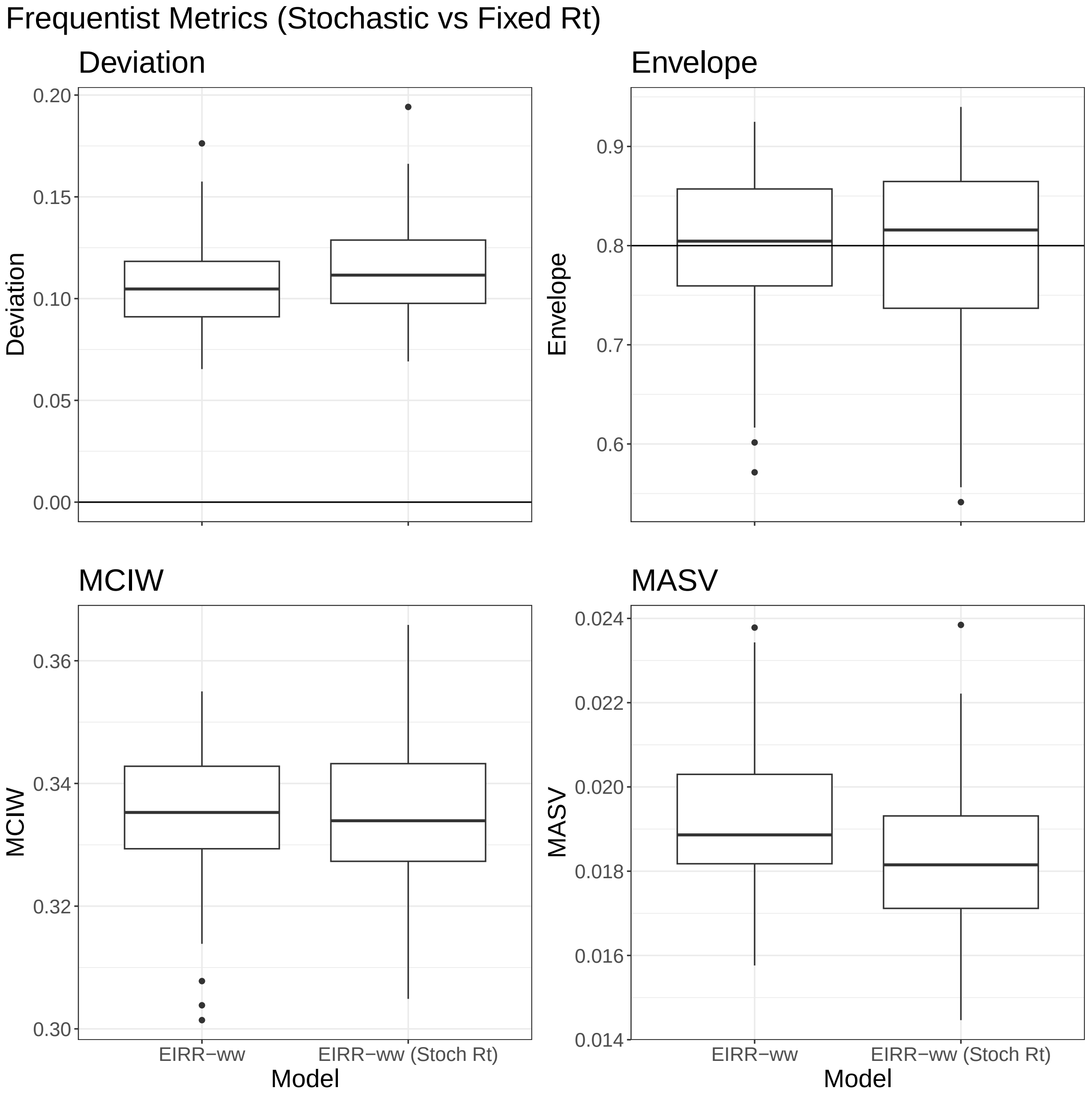


Figure B.17: Frequentistic metrics for performance of the EIRR-ww model on 100 simulations where R_t is also simulated. In the baseline scenario, R_t is fixed for all simulations, we compare to the case where R_0 is fixed, but R_t varies. See Figure 3 in the main text for descriptions of the metrics.

B.2.10 Calculating Initial Conditions for Los Angeles, CA

We used the case data to create a rough guess for the initial conditions for our EIRR-ww and EIR-cases models. We assumed the total number of individuals in Los Angeles County in the E and I compartments was equal to the last 11 days before the start of the observation period (the sum of the average latent period and average infectious period) of reported cases multiplied by 5 (i.e. an under-reporting rate of 0.2). We then split this total so that two thirds went to the I compartment and one third went to the E compartment. We then took the 18 days (the average duration in the $R1$ compartment) before the last 11 days, multiplied the total cases by 5 and assumed that this was the number of individuals in the $R1$ compartment. Los Angeles County has about 10 million people total, so we multiplied these counts by 0.48 for the final compartmnet counts, as the JWPCP plant serves 4.8 million people. For the initial effective reproduction number, we chose a prior centered around 2, this was based on previous estimates of the effective reproduction number during this time using case data [Goldstein et al., 2024]. The final priors are displayed in the table below.

Table B.4: EIRR-ww/EIR-cases Priors for Los Angeles, CA.

Parameter	Prior	Prior Median (95% Interval)
$E(0)$	Normal(2995, 0.05)	2995.00 (2994.90, 2995.10)
$I(0)$	Normal(5990, 0.05)	5990.00 (5998.90, 5990.10)
$R1(0)$	Normal(11055, 0.05)	11055.00 (11054.90, 11055.10)
R_0	Log-Normal(log(2), 0.1)	2 (1.64, 2.43)

B.2.11 Additional Estimates of R_t for SARS-CoV-2 in Los Angeles, CA

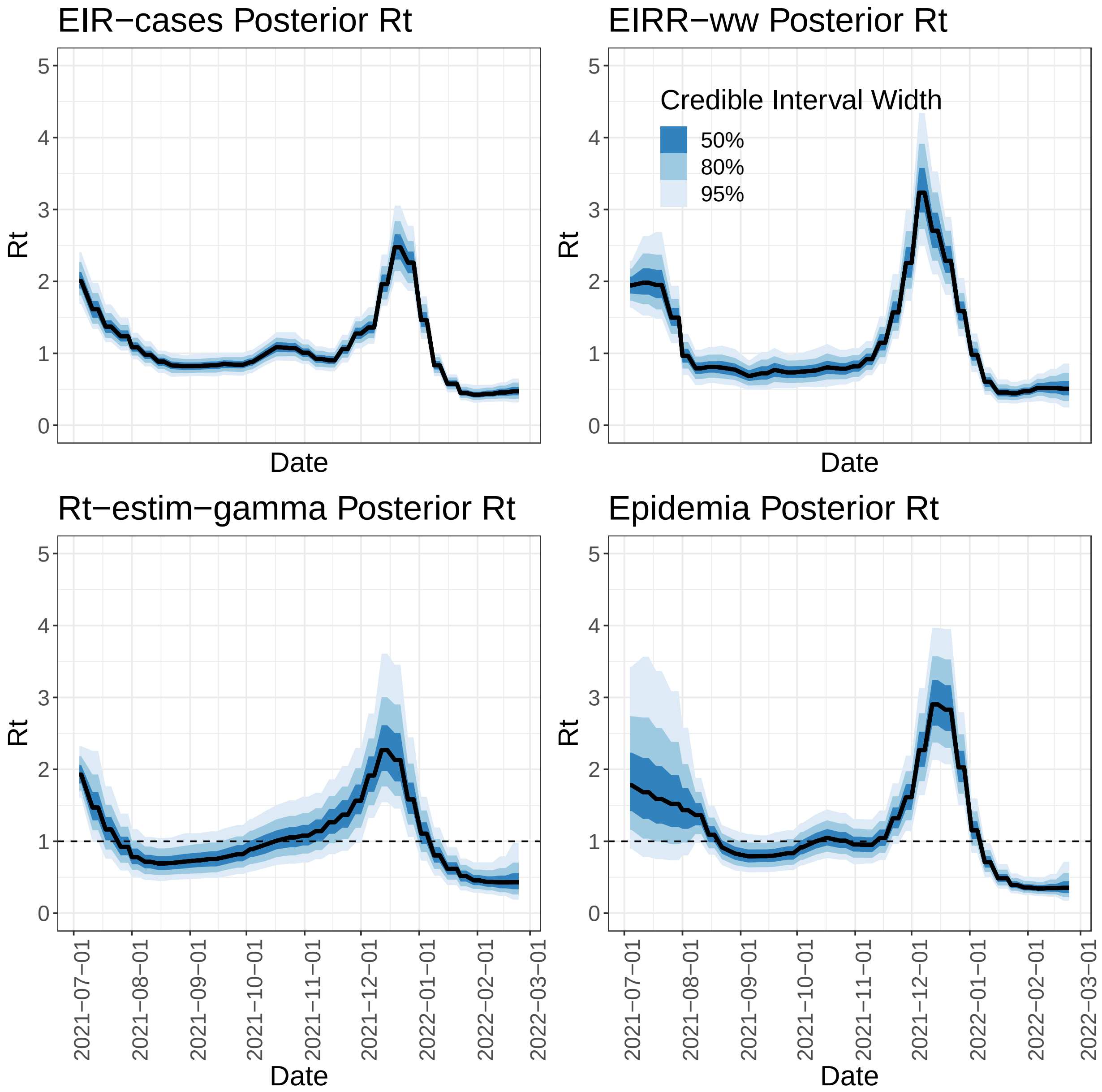


Figure B.18: Posterior estimates of the effective reproduction number for the SARS-CoV-2 epidemic in Los Angeles, CA. Blue bars from dark to light represent 50, 80, and 95% credible intervals. Black lines represent posterior medians. EIRR-ww model is fit to wastewater data. EIR-cases and epidemia are fit to cases alone. Rt-estim-gamma is fit to cases and uses total diagnostic tests as a covariate.

B.2.12 Estimates of the Case Detection Rate for SARS-CoV-2 in Los Angeles, CA

We can use the the posterior samples of $C(t_u)$ (the cumulative incidence at time t_u) produced by the EIRR-ww model to estimate the case detection rate, i.e., the proportion of new infections that are observed as reported cases. For each posterior sample, we calculate a sample of the case detection rate as

$$\kappa_u = O_u/(C(t_u - C(t_{u-1}),$$

where O_u is the number of new cases observed in the time period $(t_{u-1}, t_u]$. Note that κ_u is a function of the total diagnostic tests administered during the time period, thus, it is also interesting to look at the case detection rate normalized by the number of diagnostic tests. We call this normalized case detection rate rho, defined as

$$\epsilon_u = \kappa_u/D_u,$$

where D_u is the total number of diagnostic tests (both positive and negative) administered in the time period $(t_{u-1}, t_u]$. Posterior estimates of the case detection rate and normalized case detection rate of SARS-CoV-2 from Los Angeles, CA are shown in the figure below.

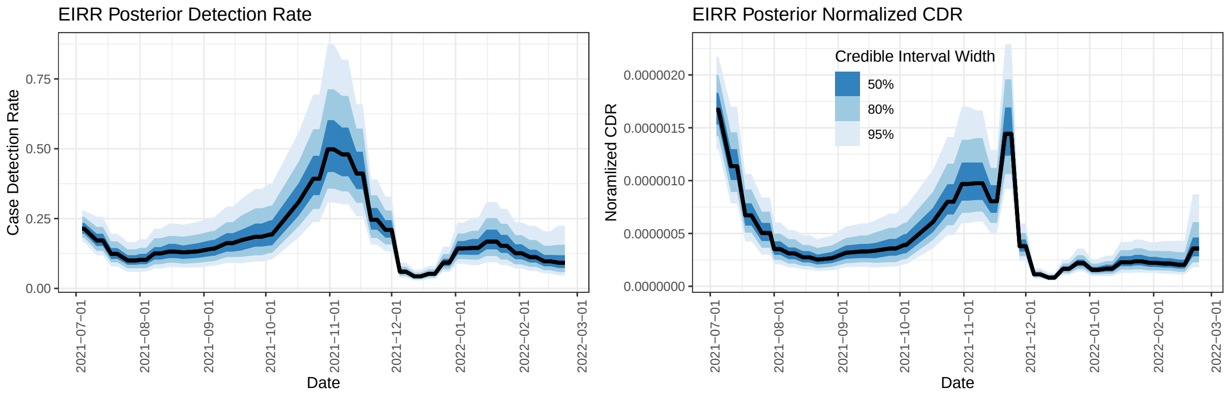


Figure B.19: Posterior summaries of the case detection rate and case detection rate normalized by total diagnostic tests for SARS-CoV-2 in Los Angeles, CA. Posterior summaries of the number of new infections taken from wastewater, and the observed weekly case counts, are used to create posterior summaries of the case detection rate.

B.3 Discussion

B.3.1 Discussion of Disagreements Between Case and Wastewater Models

An interesting point of disagreement is in the period from October 2021 to November 2021, when all three case models have a posterior median around one by the middle of October. In contrast, the EIRR-ww posterior median is well below one. One explanation for the discrepancy is that the case detection rate may have changed in October 2021. To explore this possibility, we estimated the case detection rate using posterior estimates of the incidence from the EIRR-ww model (Figure B.19), which indeed show a sharp increase in the case detection rate (both the raw rate and the rate normalized by the total diagnostic tests) in October. These estimates should be viewed skeptically, as we found that in simulation, the EIRR-ww estimates for incidence were not particularly accurate (Figure B.9). Still, we would expect the EIRR-ww model to perform worst at peaks, when the linearity assumption

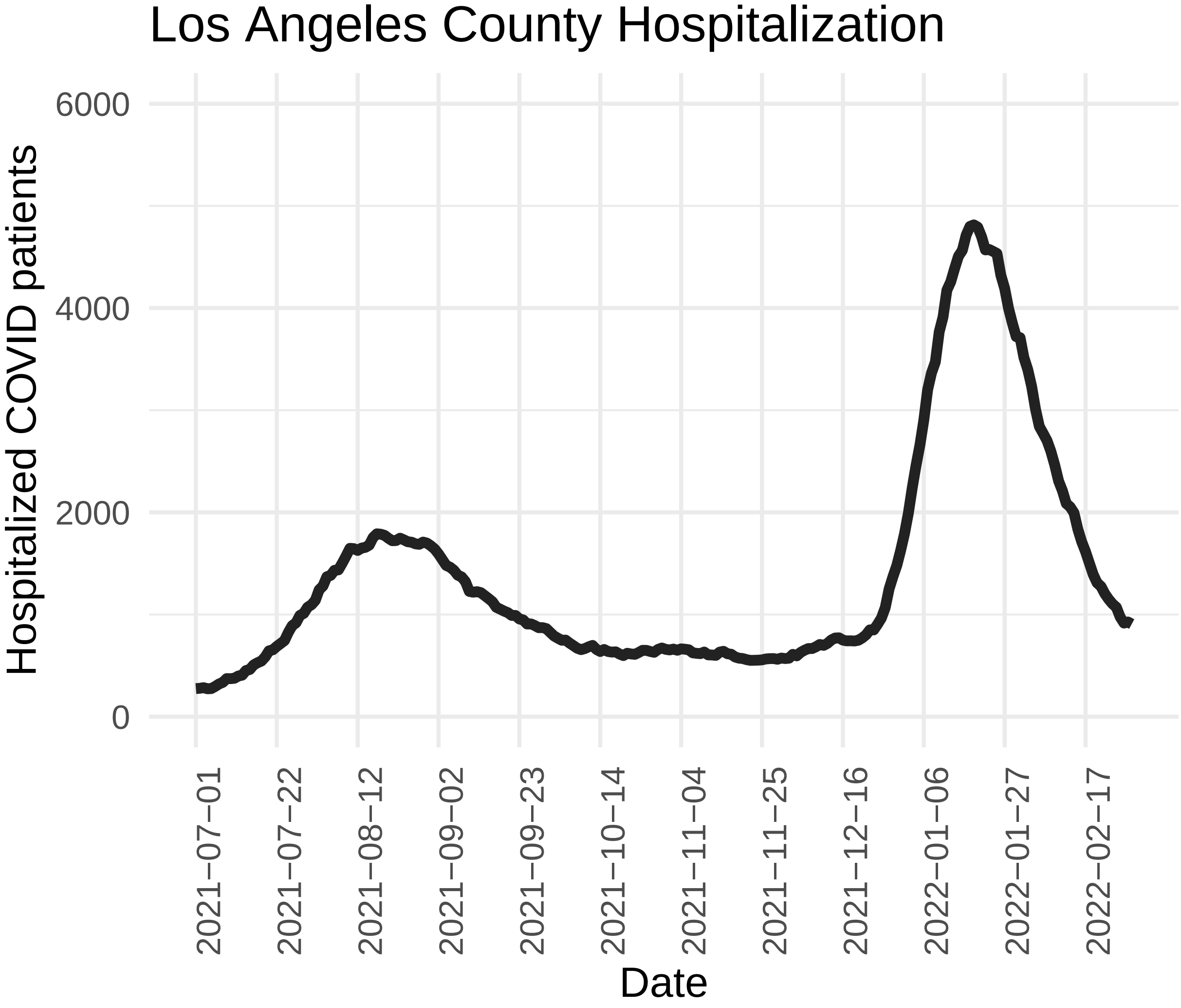


Figure B.20: Time series of patients hospitalized with SARS-CoV-2 for Los Angeles County, CA.

is particularly violated (the same counts of individuals shed different amounts of genomes if they are recently infected versus near the end of infectiousness, leading to different concentrations), so the change shown in our model estimates may still be real. When we examined the time series of hospitalizations from SARS-CoV-2 during this time period (Figure B.20), a much more reliable data source than cases, it showed no sign of an increase in transmission rate in October (which we would expect to see reflected in an increase in hospitalizations in late October/early November). The other major point of disagreement is in July 2021, when the EIRR-ww model estimates R_t to be stable around 2 until August, while every case model estimates a steady decline in R_t. Here again the change in case detection rate may be to blame, but the hospitalizations are a little harder to interpret. The flatter hospitalization curve in August of 2021 may indicate the EIRR-ww model is correct, on the other hand, if R_t was indeed still at 2 on August 1st, we might expect a peak of hospitalizations even later than mid-August. We are inclined to think the case models are more correct in this case.

Appendix C

Additional Material for Chapter 5

C.1 EI Moment ODEs

Recall $\mathbf{H}(t) = \{E(t), I(t)\}$ is a Markov Jump Process where $E(t)$ is the number of infected but not yet infectious individuals at time t, and $I(t)$ is the number of infectious individuals at time t. To simplify notation, we write $E[E] = E[E(t)|E(0) = e_0, I(0) = i_0]$. Then, the changes in the conditional first and second moments of $\mathbf{H}(t)$ are described by the following ordinary differential equations:

$$\begin{aligned}
\frac{dE[E]}{dt} &= \alpha E[I] - \gamma E[E], \\
\frac{dE[I]}{dt} &= \gamma E[E] - \nu E[I], \\
\frac{d[E^2]}{dt} &= \alpha E[I] + \gamma E[E] + 2\alpha E[EI] - 2\gamma E[E^2], \\
\frac{dE[I^2]}{dt} &= \nu E[I] - 2\nu E[I^2] + \gamma E[E] + 2\gamma E[EI], \\
\frac{dE[EI]}{dt} &= \alpha E[I^2] - \nu E[EI] - \gamma E[E] + \gamma E[E^2] - \gamma E[EI].
\end{aligned}$$

Using the definition of variance and covariance, we arrive at ODEs for the conditional variance and covariance as well:

$$\begin{aligned}
\frac{d\text{Var}(E)}{dt} &= \alpha E[I] - 2\gamma \text{Var}(E) + \gamma E[E] + 2\alpha \text{Cov}(E, I), \\
\frac{d\text{Var}(I)}{dt} &= \nu E[I] - 2\nu \text{Var}(I) + \gamma E[E] + 2\gamma \text{Cov}(E, I), \\
\frac{d\text{Cov}(E, I)}{dt} &= \alpha \text{Var}(I) + \gamma \text{Var}(E) - \gamma E[E] - (\nu + \gamma)\text{Cov}(E, I).
\end{aligned}$$

C.2 Simulation Comparison of MJP vs Log-Normal

To visualize the difference between our Log-Normal process and the true Markov jump process, we simulated 1000 data sets from each model using a fixed value of $R_t = 1.5$ with 10, 20, and 30 individuals starting in the E and I compartments. We simulated from the MJP using the classic Gillespie algorithm [Gillespie, 1977]. Marginal quantiles of the counts in each compartment at times 1 through 30 are displayed in Figure C.1.

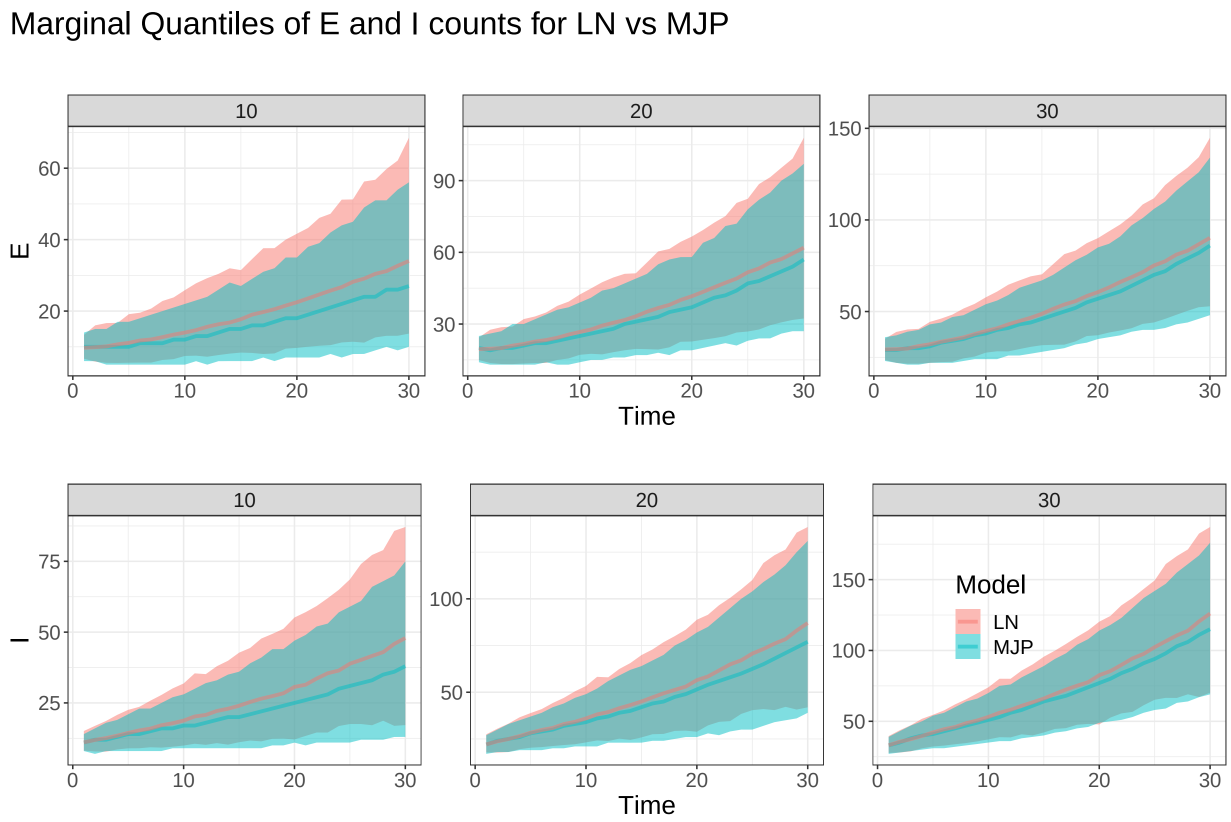


Figure C.1: Empirical marginal quantiles of the counts in the E and I compartments for the original MJP EI model and the LN approximate model simulated from 1000 data sets. The lines are medians, the ends of the shaded regions are the 2.5% and 97.5% quantiles. Each panel shows the distribution with a different number of individuals in the E and I compartments at the start of the simulation, all other parameters in the models are identical. As the number of individuals increases, the two distributions look more similar.

As we would expect, as the initial number of individuals in the simulation increases, the distributions of the MJP and Log-Normal approximate model look increasingly similar [Barbour, 1974].

C.3 Obvservation Models using t-Distribution versus Normal Distribution

As stated in the main text, we chose to use a simpler observation model than the one used in Chapter 4. The t-distribution has fatter tails than the normal distribution, however the stochastic model has a latent stochastic epidemic process, which should accommodate additional levels of noise as well. Additionally, we expect that in the small populations, the wastewater data may actually be less prone to outliers than in large populations, simply because the data is collected much closer to the source in small population settings. That is, we expect less degradation and transformation due to time spent in the sewer system, simply because the journey from toilet to sampler is shorter in a small population setting. As additional evidence that the normal distribution is adequate, we include the posterior predictive summaries of our models fit to real world data, which show no signs that the model is unable to simulate real data (Figure C.4). There is much room to explore this topic in future efforts.

C.4 Epidemia-cases

The Epidemia-cases model relies on the so-called renewal equation which is described in detail in Chapter 3.

The `epidemia` package can be used to create different branching process inspired models to estimate the effective reproduction number using different observation models and models for latent incidence [Bhatt et al., 2023]. For the model we used in this chapter, we modeled observed cases using a negative binomial distribution, modeled the effective reproduction number as a Gaussian random walk, and modeled unobserved incidence as an auto-regressive normal random variable with variance equal to the mean multiplied by an over-dispersion parameter. In additional, we modeled the case detection rate as a change point model with a baseline value plus an indicator function with a covariate, reflecting the change in UCI testing policy that occurred in March 2022. The explicit model is listed below:

$$\begin{aligned}
\tau &\sim \exp(\lambda)\text{–Hyperprior for unobserved incidence,}\\
I_\nu &\sim \exp(\tau)\text{–Prior on unobserved incidence } \nu \text{ days before observation,}\\
I_{\nu+1},\dots,I_0 &= I_\nu\text{–Unobserved incidence,}\\
\sigma &\sim \text{Truncated-Normal}(0,0.15)\text{–Prior on variance of random walk}\\
\log R_0 &\sim \text{Normal}(\log 0.5, 0.1)\text{–Prior on } R_0,\\
\log R_t|\log R_{t-1} &\sim \text{Normal}(\log R_{t-1},\sigma)\text{–Random walk prior on } R_t,\\
\psi &\sim \text{Normal}(10,2)\text{–Prior on variance parameter for incidence,}\\
I_t|I_\nu,\dots,I_{t-1},R_t &\sim \text{Normal}(R_t\sum_{s<t} I_s g_{t-s},\psi)\text{–Model for incidence,}\\
\text{logit}\alpha &= \beta_0+\beta_1\text{Policy Change–Case detection rate model,}\\
\beta_0 &\sim \text{Normal}(-1.1,0.2)\text{–Baseline case detection rate,}\\
\beta_1 &\sim \text{Normal}(0,0.5)\text{–Difference in case detection post policy change,}\\
y_t &= \alpha\sum_{s<t} I_s\pi_{t-s}\text{–Mean of observed data model,}\\
\phi &\sim P(\phi)\text{–Prior on dispersion parameter for observed data,}\\
Y_t &\sim \text{Neg-Binom}(y_t,\phi)\text{–Observed data model.}
\end{aligned}$$

Here π_t are the values of the probability density function for the delay distribution, the time between an individual being infected and being observed.

C.5 Simulation Protocol

C.5.1 Simulation Engine

We used the same simulation engine and parameters described in Appendix Section B.2.1 and B.2.3. The only difference was that the population was set to be 1000, with 10 individuals starting in the E and I compartments. In addition, we simulated the times at which the initial individuals became infectious or became infected so that the individuals were all shedding different pathogen genome concenetrations at the start of the simulation. We simulated these additional times from the corresponding distributions of the SEIRR, so the time individuals became infectious were simulated from an exponential distribution with parameter ν. We then set the time becoming infectious as a negative value, so it happened before the start of the simulation at time 0. These times were simulated after the epidemic itself was simulated, so they did not affect the dynamics of the simulation itself, only the concentration of RNA being shed.

C.5.2 Simulation Priors

The priors used in the steep simulation scenario are listed

Table C.1: Priors used by all models in the steep simulation scenario.

Parameter	Model	Prior	Prior Median (95% Interval)	Truth
γ	All	Log-normal(log(1/4), 0.2)	0.25 (0.17, 0.37)	0.25
ν	All	Log-normal(log(1/7), 0.2)	0.14 (0.10, 0.21)	0.14
σ_{rw}	All	Log-normal(log(0.15), 0.2)	0.1 (0.10, 0.22)	NA
τ	All	Log-normal(0, 1)	1.00 (0.14, 7.10)	0.5
ρ	All	Log-normal(0, 1)	1.00 (0.14, 7.10)	NA
R_0	All	Log-Normal(log(0.8), 0.1)	0.8 (0.66, 0.97)	0.80
$E(0)$	All	Log-Normal(log(10), 0.05)	10 (9.07, 11.03)	10
$I(0)$	All	Log-Normal(log(10), 0.05)	10 (9.07, 11.03)	10

For the shallow scenario, we centered the prior on 0.9 rather than 0.8, for the fixed scenarios, we centered the prior at 1.3, but the standard deviation parameter remained the same. Similarly, the priors on initial compartment accounts were shifted to 5 or 20 for the initial5 and initial20 scenarios. Otherwise, priors remained the same across scenarios.

C.6 Priors and Posteriors of Fixed Parameters

We provide a visualization of the priors and posteriors of the fixed parameters in the stochastic EI-ww model for the model fit to the Fixed scenario visualized in Figure 5.2. We also show the prior and posterior of τ in particular, as the scale is hard to see in the original figure.

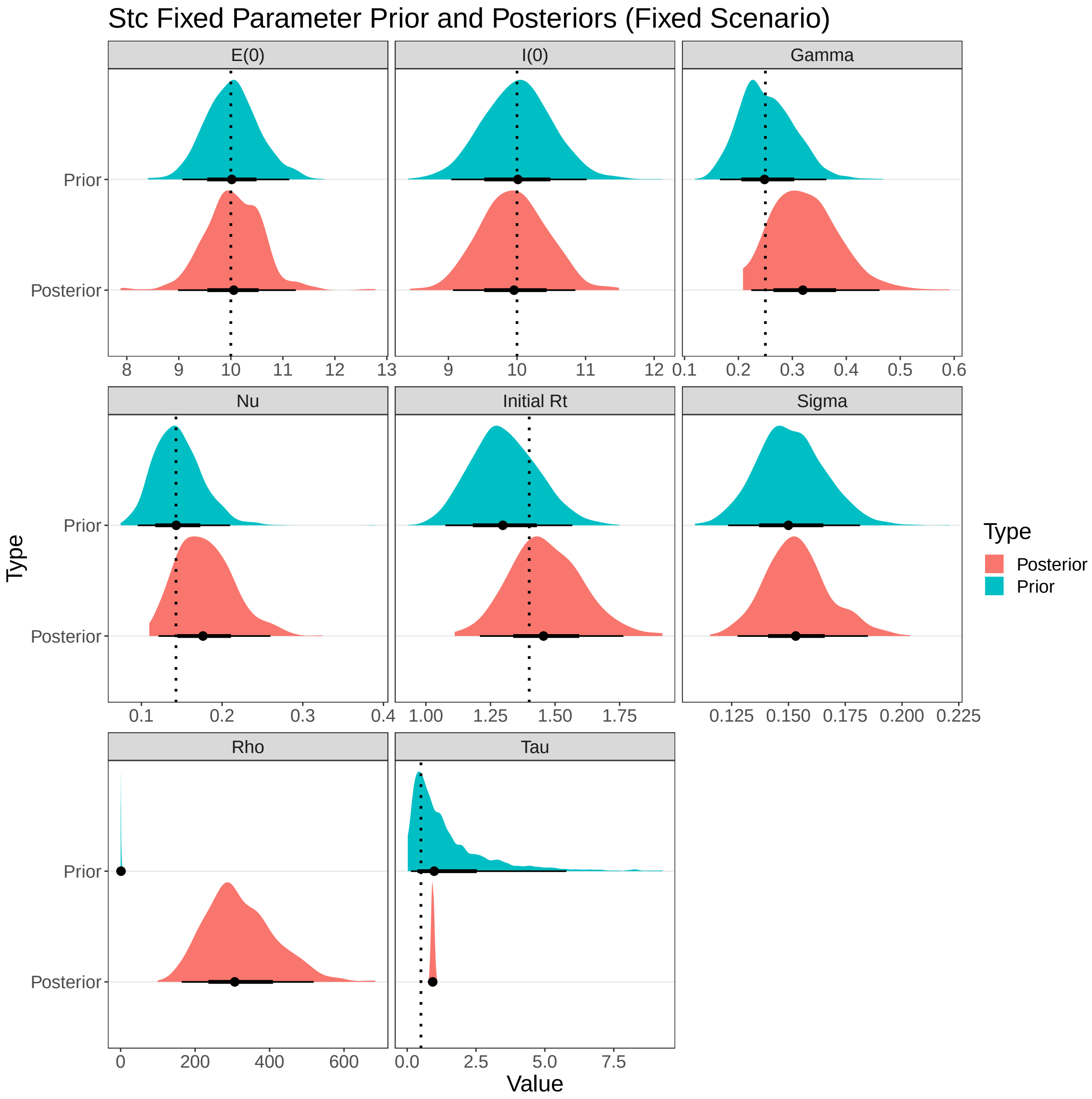


Figure C.2: Stochastic EI-ww prior and posterior densities for fixed model parameters. Posterior summaries are from the model fit to the Fixed Scenario data shown in Figure B.9. Blue densities are the prior, red densities are the posterior, dotted lines indicate true values (when relevant).

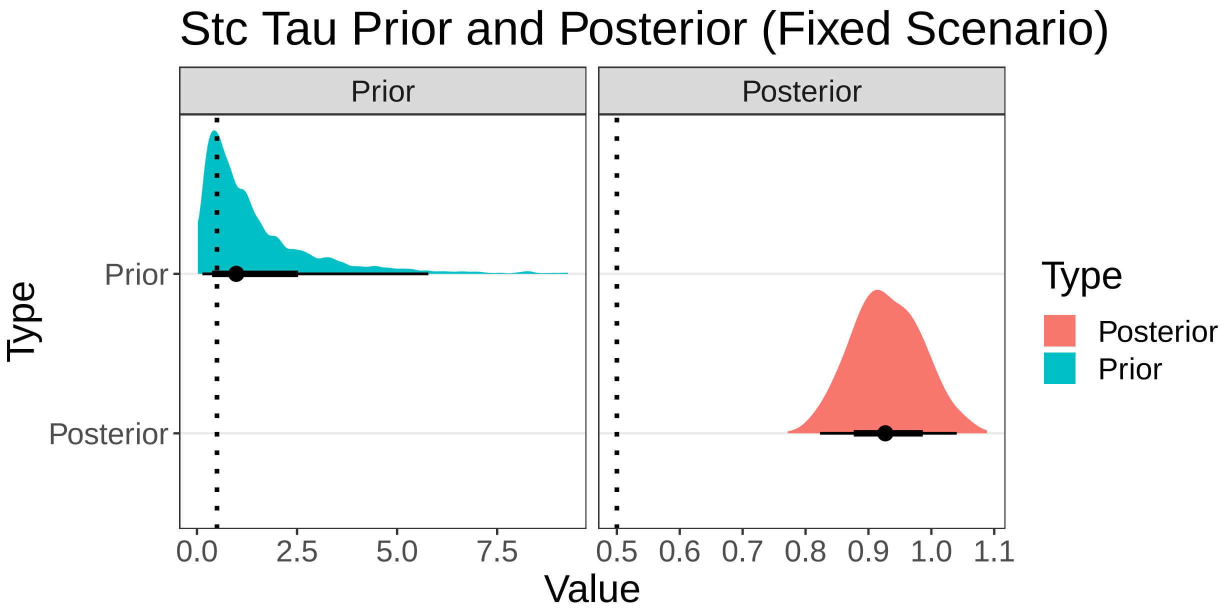


Figure C.3: Stochastic EI-ww prior and posterior densities for fixed model parameters for the parameter τ. The figure is a transformation of the bottom right panel of Figure C.2.

C.7 Calculating Initial Conditions for Real Data

We used case data to create a rough guess for the initial conditions for our stochastic and deterministic EI-ww models. We assumed the total number of individuals in the E and I compartments was equal to the last 11 days before the start of the observation period (the sum of the average latent period and average infectious period) of reported cases multiplied by 4 (i.e., a case detection rate of 0.25). We then split this total so that two thirds are allocated to the I compartment and one third went to the E compartment. These case data were available at the community level, we then split the calculate counts proportionally by the proportion of the community living in each sub-community. For the initial effective reproduction number, we chose a prior centered around 0.5, this was based on previous estimates of the effective reproduction number during this time using case data (Chapter 4, Goldstein [2024]). The final priors are displayed in the table below.

Table C.2: EI-ww priors for UC Irvine.

Parameter	Place	Prior	Prior Median (95% Interval)
$E(0)$	E1	Log-Normal(log(5), 0.05)	5 (4.53, 5.52)
$I(0)$	E1	Log-Normal(log(9), 0.05)	9 (3.63, 4.41)
$E(0)$	G1	Log-Normal(log(6), 0.05)	6 (5.44, 6.62)
$I(0)$	G1	Log-Normal(log(10), 0.05)	10 (9.07, 11.03)
$E(0)$	G2	Log-Normal(log(7), 0.05)	7 (6.35, 7.72)
$I(0)$	G2	Log-Normal(log(12), 0.05)	12 (10.88, 13.24)
R_0	Log-Normal(log(0.5), 0.1)	0.5 (0.41, 0.61)	

C.8 Posterior Predictive Real Data Analysis

C.9 Sensitivity Real Data Analysis

In the main analysis we used an initial R_t derived from previous case-based estimates of R_t for the entirety of Orange County, CA. It is plausible this county-wide R_t did not accurately reflect the situation at UC Irvine at the start of the modeling period. To understand how our choice of prior affected our analysis, we re-analyzed the data using an alternative prior centered around 1 instead of 0.5 ($R_0 \sim \text{Log-Normal}(0, 0.1)$). Figure C.5 displays the results of this re-analysis.

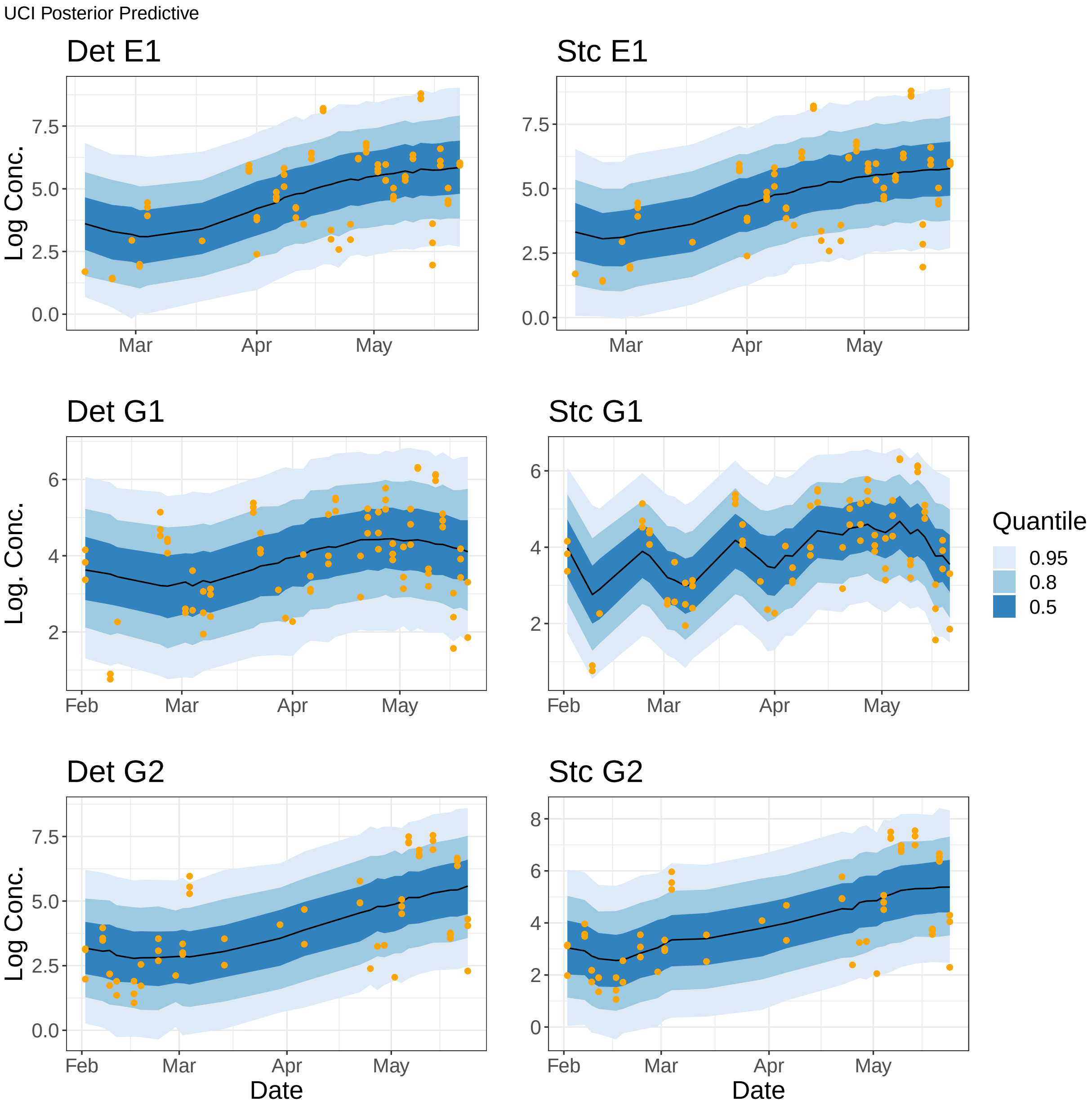


Figure C.4: Posterior predictive summaries from UCI wastewater data. Real data are orange dots, the posterior predictive median is the black line, blue regions are distribution quantiles of varying width.

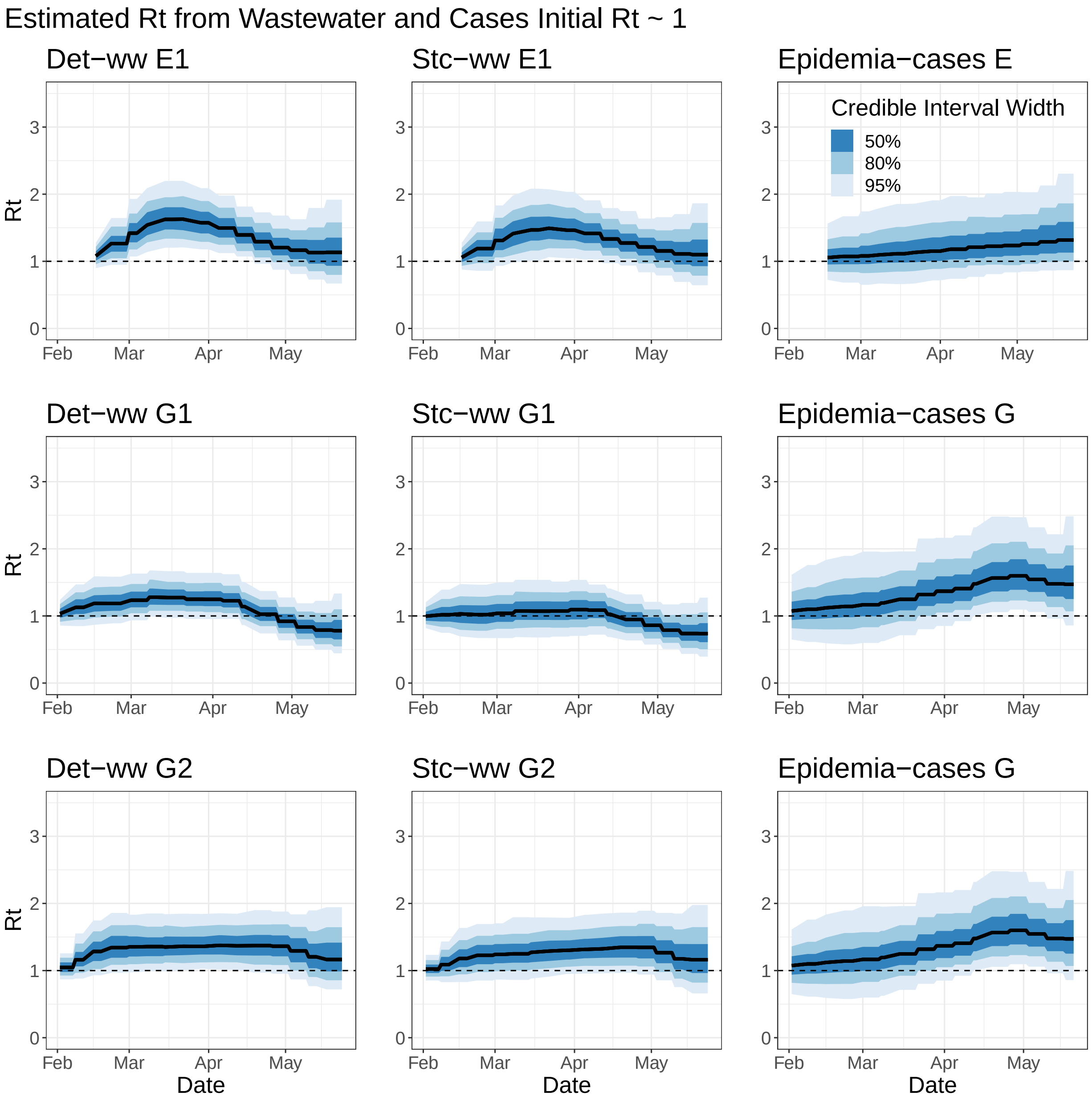


Figure C.5: Posterior summaries of the effective reproduction number in college campus communities estimated from wastewater data and case data using a prior on the initial R_t centered at 1 rather than 0.5. Black lines are medians, blue shaded regions are credible intervals.

Many of the patterns we observed in the main analysis are still present. The wastewater model results are quite different from case model results. The deterministic EI-ww model is a little less certain than in the main analysis, but still more certain overall than the stochastic

EI-ww model. In the case of the E community, the stochastic model now has 95% credible interval above 1 for a brief time period, while the case model never has 95% credible intervals above 1.