Title
On Recursive and Hawkes Models for Forecasting the Spread of Epidemic Diseases

Author
Kaplan, Andrew

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

On Recursive and Hawkes Models for Forecasting the Spread of Epidemic Diseases

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Andrew Marcus Kaplan

2022

ABSTRACT OF THE DISSERTATION

On Recursive and Hawkes Models for Forecasting
the Spread of Epidemic Diseases

by

Andrew Marcus Kaplan
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2022
Professor Frederic Schoenberg, Chair

The self-exciting Hawkes point process model (Hawkes, 1971) has been used to describe and forecast communicable diseases. In this dissertation, there are two parts. First, we introduce the non-parametric version of the recursive model (Schoenberg et al., 2019), an adaptation of the Hawkes model which allows for variable productivity, or disease reproduction rate. Here, we extend the data-driven non-parametric EM method of Marsan and Lengliné (2008) in order to fit the recursive model without assuming a particular functional form for the productivity. We then evaluate the ability of the non-parametric recursive model to fit and forecast cases of mumps in Pennsylvania compared to that of other point process models and a variation of the commonly used SIR (Susceptible, Infected, Recovered) compartmental model. Second, we examine increasing surges of the COVID-19 pandemic using the HawkesN model with an exponential kernel (Rizoiu et al., 2018), which assumes a finite susceptible population, is considered stationary when the reproduction number κ is greater than one, and has interpretable terms similar to that of the SIR model (Kresin et al., 2021). The HawkesN model is fit using a least squares method introduced in Schoenberg (2021), which

is an effective method when an epidemiologic dataset only provides a daily case count rather than a specific time of infection. We first examine doubling time of COVID-19 in California during three notable surges using the HawkesN model and compare its predictive ability to that of an adaptation of the SIR compartmental model. Secondly, we compare HawkesN to the same compartmental model in forecasting cases of SARS-COV-2 for all fifty states nationwide. This larger study is to guide further work in improving the predictive ability of HawkesN.

The dissertation of Andrew Marcus Kaplan is approved.

Karen McKinnon

Ying Nian Wu

Mark Handcock

Frederic Schoenberg, Committee Chair

University of California, Los Angeles

2022

To the millions of people

who perished

or fell on hard times

due to the COVID-19 pandemic

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It has been a real adventure and rare opportunity not only to complete the research necessary to write this dissertation, but also to do it during a once in a lifetime COVID-19 pandemic. While I may not have gone to as many conferences as some aspiring researchers do, I have had the privilege of furthering the science of disease modeling during an active pandemic which has brought an extra sense of importance and urgency to my research. It has also enhanced the meaning of my day-to-day tasks. However, getting to this point would not be possible without the help and guiding hand of many.

First, I would like to thank my advisor Dr. Frederic Schoenberg for all of his help guiding me through the Ph.D. process. We first crossed paths when I took his course in point process methods during the first year of the Ph.D. program and I found myself quite interested in this area. It happened that he was just beginning some of his work in applying point process models to epidemiologic processes at the time and I thought that my background and interest in biostatistics would be a great fit for this research niche. This turned out to be the case, and years later, here I am. Professor Schoenberg has not only been instrumental in getting me to this point, he has helped me to further develop my skills as a more active, independent researcher who strives to keep up with the latest developments in the field at large, while staying focused on my own area of study. Whenever I felt like I was stuck in some aspect of my research and had no way out, he was always on hand with insightful suggestions and moral support to help me get back on track.

I would also like to thank the rest of my committee including Dr. Mark Handcock, Dr. Ying Nian Wu and Dr. Karen McKinnon for their involvement and advice as well. They were also instrumental in helping me sharpen my sills as a researcher and their doors were always open when I needed any extra guidance. In addition, classes they taught including those on machine learning, networks and monte carlo simulation were excellent and added

significantly to my statistical knowledge base.

Besides my advisor and committee members, I would also like to thank all of the outstanding lecturers in the department. I have had the privilege of being a teaching assistant for many of them including Linda Zanontian, Guani Wu, Mike Tsiang, Dave Zes, Bingling Wang, Akram Almohalwas, Maria Cha and Vivian Lew. All have been great sources of support over the years and helped me to hone my own teaching skills as a TA which I found very rewarding.

One name I haven't mentioned yet is that of Nicolas Christou. He has been supportive on a whole other level. He always made himself readily available to me and was often the first person besides my advisor I would turn to for his guidance and insights. Also, he is well known for being an outstanding teacher and I couldn't agree more. His geostatistics class was one of my favorites.

I would also like to acknowledge the staff at the Statistics Department for making everything from enrollment to student employment easy. This included the Student Affairs Officer Glenda Jones, Laurie Leyden and Chie Ryu among others. To be honest, whenever GSA reps would call me asking my opinion on how I was treated, funding and housing issues, etc. I would always answer that there were no issues. The department was always conscientious about taking care of these items every quarter, which was quite helpful to say the least.

Most importantly, I would like to thank my family for all of their support during this long and arduous journey. My parents Susie and Jon Kaplan have always been there to encourage me along the way, especially when the going got tough. I also really appreciate their willingness to accommodate me during the long period of isolation in the midst of the acute period of the COVID-19 pandemic. I also want to thank my brother Sam, his partner Rhina

and my niece Charlotte for their love and guidance. Others who have been critical to my success are family members, Melinda Price & Chanse Moody, Kathy, Dan & Lily Yarmo as well as Ken, Stephanie, Leo & Ari Price. If it wasn't for all of your love and support over these years, I would never have gotten here.

Last, but not least, I would like to acknowledge the outside sources critical to my research. The material presented in Chapter 2 is a version of the manuscript: Kaplan AM, Park J, Kresin C, and Schoenberg FP. (2021). Nonparametric estimation of recursive point processes with application to mumps in Pennsylvania. *Biometrical Journal*, DOI: https://doi.org/10.1002/bimj.202000245. It also includes the open data badge for Reproducible Research awarded by the *Biometrical Journal*. It is based upon work supported by the National Science Foundation under grant number DMS 1513657. Thanks to the US CDC for supplying the data and to Project Tycho Van Panhuis et al. (2018) for making it so easily available.

The information presented in Chapters 3 & 4 is based upon work supported by the National Science Foundation under grant number DMS-2124433. We also thank the CDC and California Department of Public Health for providing the data for our analysis. For all chapters, computations were performed in R (https://www.r-project.org).

VITA

2006-2010	B.S. Statistics with minors in Math and Psychology, California Polytechnic State University, San Luis Obispo
2010-2011	M.A. Applied Statistics, University of California, Santa Barbara
2012-2017	Statistician, Department of Radiology, University of California, Los Angeles
2017-2020	Statistical Consultant, MedQIA, Los Angeles, CA
2018-present	Teaching Assistant/Associate/Fellow Statistics Department, UCLA. Taught sections of Introduction to Statistics, Introduction to Statistical Programming with R, Introduction to Mathematical Statistics, Introduction to Design and Analysis of Experiments, Introduction to Monte Carlo Methods & Intro to Databases (MAS program)
2018-2020	Graduate Student Researcher, Statistics Department, UCLA. Performed point process analysis on the effects of crack propagation within bullet-proof armor when struck. Processed 3D DICOM images (sliced) into R, performing statistical analyses and providing periodic updates.
2022	Biostatistics Intern, Janssen Statistics and Decision Sciences

PUBLICATIONS

Kaplan AM, Park J, Kresin C, and Schoenberg FP. (2021). Nonparametric estimation of recursive point processes with application to mumps in Pennsylvania. *Biometrical Journal*,

DOI: https://doi.org/10.1002/bimj.202000245. Note: includes the open data badge for Reproducible Research.

Kaplan A, Kresin C, and Schoenberg FP. (2022, in progress) Estimation of doubling time for SARS-COV-2 in California using HawkesN and SQUIDER models.

Kaplan A, Farzin-Nia S, and Schoenberg FP. (2022, in progress) Evaluation of HawkesN model using least squares fitting method in estimating SARS-COV-2 in the United States.

CHAPTER 1

Introduction

1.1 Point Processes

Point process modeling is a well established technique used to model sequences of outbreaks or event aftershocks which occur at random time intervals or locations (Schoenberg, 2017). A point process is a collection of events that occur in a defined space such as earthquakes, lightning strikes, wildfires, or disease outbreaks. A temporal point process can be described as a counting method where for any time t in the timeframe $[0, T]$, $N(t)$ represents the number of points that occurred at or before t (Schoenberg, 2017). In other words, a temporal point process model takes into account the number of events happening in a particular time frame as well as any clustering of those times. This is contrary or in addition to a spatial point process model which describes any possible clustering or inhibition of points spatially within a defined region.

1.2 Self-Exciting Point Processes

The self-exciting Hawkes point process model (Hawkes, 1971) is commonly used to describe temporal clustering phenomena. The self-exciting component refers to the fact that any prior event can trigger future ones. The strictly temporal version of the Hawkes model for the conditional intensity function λ is defined over the time frame $[0, T]$ as

$$\lambda(t) = \mu + \kappa \int_0^t g(t - t^{'}) dNt^{'} \tag{1.1}$$

In other words, the conditional intensity, or the expected number of events at a given time t, is only dependent on the background rate μ, the productivity κ, and the triggering function g. The triggering function g is a density function which estimates the amount of self-excitation caused by previous events and the κ term is assumed constant in a simple Hawkes model (Hawkes, 1971). It is used in a wide variety of applications ranging from modeling earthquake patterns (Ogata, 1988), (Ogata, 1998), to crime sprees (Mohler et al., 2011), (Park et al., 2021) and to the spread of invasive species (Balderama et al., 2012).

While the initial use of the Hawkes model has been to model earthquakes (Ogata, 1988), (Ogata, 1998), expanded versions of the Hawkes models have been applied to communicable disease processes more recently. In epidemiologic settings, the rapid spread of a disease from one host to another can be taken into account using the self-excitation property of the Hawkes model. One of the first applications of the Hawkes model to disease processes is discussed in Meyer et al. (2012) where meningococcal outbreaks in Germany have been successfully described using a spatial-temporal model with an endemic and an epidemic component (Meyer et al., 2017). The endemic portion represents the background rate and the epidemic part allows for any prior infection to trigger future events; the self-exciting component. The triggering component also depends on the intensity of a particular outbreak (Meyer et al., 2017). Another successful application is demonstrated in Park et al. (2022) in which the spread of Ebola in West Africa is predicted using a non-parametric representation of the Hawkes model. It is mentioned in more detail later.

1.3 Purpose

Even though the simple Hawkes model has performed well in modeling several communicable disease processes, its application is still relatively new and leaves substantial room for improvement. For instance, the simple Hawkes model is not a stable (stationary) process

whenever there is an increase in cases, or $\kappa > 1$ (Park et al., 2022). With κ assumed constant, further triggering implies that an infection will result in an expected $\kappa+\kappa^2+\kappa^3+... = \frac{1}{1-\kappa}-1$ new cases, meaning the process is only considered stationary if $0 < \kappa < 1$. Also, the notion that the productivity κ should be held constant, as assumed in the simple Hawkes model, is questionable. It is mentioned in Schoenberg et al. (2019) that in purely temporal epidemiological point process modeling, the expected number of transmissions for a subject infected at time t should depend on the conditional intensity at time t. This is because when the prevalence of the disease is low early on, the rate of transmission should be much higher than it is for later recurrences when the virus has already been spread. The rationale for such an argument is that human intervention, vaccines and prior exposure to the disease would reduce morbidity (Schoenberg et al., 2019).

In this dissertation, two new adaptations of the Hawkes model, meant to address these weaknesses, are explored in detail. The recursive model, introduced in Schoenberg et al. (2019), allows for a non-constant productivity or κ. In Schoenberg et al. (2019), the recursive model is introduced in a parametric form, but here, it is shown that it is possible to create a non-parametric version of the recursive model that is more effective at fitting and forecasting cases of mumps in Pennsylvania, originally published in Kaplan et al. (2021). The first part of the dissertation introduces the non-parametric recursive model in detail.

The HawkesN model (Rizoiu et al., 2018), allows for a finite susceptible population, thus allowing for stationarity when $\kappa > 1$ (Kresin et al., 2021). Also, it is shown that the parameters in the HawkesN model can be estimated using a least squares method introduced in Schoenberg (2021), which permits an easier estimation when the exact time of day of an infection is not provided, as conventional in most epidemiologic datasets. In the second part of the dissertation, we take advantage of the newfound properties offered by the HawkesN model and examine how the various HawkesN parameters affect doubling time during three

surges of the COVID-19 pandemic in California. The third part of this work is a larger analysis where case counts for all fifty states during three COVID-19 surges are fit and forecast using both the HawkesN model as well as an adaptation of the commonly used Susceptible, Infected, Recovered (SIR) compartmental model. The goal here is to determine specific instances when HawkesN performs better than the compartmental model or vice-versa in order to guide future work.

CHAPTER 2

Non-Parametric Recursive Model

2.1 Introduction

As described in Chapter 1, the simple Hawkes model makes the assumption that the number of "aftershocks" triggered by a current event, known as productivity, is constant (Schoenberg et al., 2019). However, it is pointed out in Schoenberg et al. (2019) that a non-constant productivity is warranted to take into account various patterns often seen in disease processes. For instance, early in an outbreak, when the prevalence of the disease is low, the rate of transmission may be much higher than at later times when the virus has already spread, due to differences in awareness, human mitigation efforts, and prior exposure of the population to the disease. In Schoenberg et al. (2019), the parametric version of the recursive model is written as

$$\lambda_t = \mu + \int_0^t H(\lambda_{t'}) g(t - t^{'}) dN_t^{'} \tag{2.1}$$

As in the simple Hawkes model, the conditional intensity $\lambda(t)$ is dependent on the background rate and the stepwise function g which takes into account the time lags between events. However, the conditional intensity in the recursive model is also a function of H, another stepwise function which depends on the rate (conditional intensity) of previous events. In other words, H is the function which determines the productivity for a subject afflicted at time t. Schoenberg et al. (2019) compares the fit of the recursive model to that of the simple Hawkes model in describing known cases of Rocky Mountain Spotted Fever in California between 1960 and 2011. The log-likelihood for the recursive model was higher than that of the

Hawkes model and the AIC criterion was lower by a statistically significant margin (Schoenberg et al., 2019). Despite the added complexity to the model, the recursive component is shown beneficial in forecasting the disease over time.

While the findings from Schoenberg et al. (2019) are significant, both the Hawkes and recursive models are estimated parametrically using maximum likelihood estimation, which assumes a specific pattern for the g and h functions. Here, we develop a non-parametric model which allows for more flexibility when fitting it to data, which in turn results in more accurate forecasts. The idea expands upon the technique used in Kelly et al. (2019), Park et al. (2022) where a non-parametric representation of the simple Hawkes model outperformed the traditional SEIR (Susceptible, Exposed, Infected, Recovered) model in predicting the spread of Ebola in West Africa. The method used most recently in Park et al. (2022) involves non-parametric optimization using an expectation maximization algorithm originally developed by Marsan and Lengliné (2008) to analyze seismic data. Described in more detail in the Background section, the E-M algorithm in Marsan and Lengliné (2008) assumes a stepwise triggering function, thus allowing for estimation of the step heights as though they are parameters. In addition, non-parametric Hawkes processes have been extended to applications in other areas such as renewal immigration (Wheatley et al., 2014), online learning algorithms (Yang et al., 2017), finance (Kirchner and Bercher, 2018) and in estimating civilian fatalities in the Iraq War (Lewis and Mohler, 2011).

The non-parametric recursive model is evaluated using epidemiological data by comparing the accuracy of both the model fit and forecast to those of other point process models, as well as more widely used compartmental models such as the SEIR (Susceptible, Exposed, Infected, Recovered) compartmental model first introduced by Kermack and McKendrick (1927) and specifically its extension, the SVEILR model introduced by Li et al. (2018). The SVEILR model is a system of differential equations that can be adjusted to account

for varying methods of transmission and rates of exposure, vaccination and recovery. Of particular interest here is that the proposed non-parametric version of the recursive model can outperform both the Hawkes and SVEILR models in its ability to fit epidemic data and forecast future cases. We compare the models using 32 years of reported cases of mumps in Pennsylvania, fitting using training data from January 1970 to September 1990 and then assess their fit using data from October 1990 to December 2001. This chapter provides a more detailed version of the Kaplan et al. (2021) manuscript, published in the Biometrical Journal with a reproducibility badge as of August, 2021.

2.2 Background

Parametric Hawkes models are conventionally estimated by maximum likelihood estimation (MLE) (e.g. (Ogata, 1988)), and the resulting estimates have desirable asymptotic properties such as consistency, asymptotic normality and efficiency (Ogata, 1978). However, in practice often the likelihood function is quite flat around its maximum, and optimization methods may fail to converge or depend greatly on the choice of starting values (Schoenberg, 2013). As an alternative, Veen and Schoenberg (2008) suggested a method based on the expectation-maximization (EM) algorithm to approximate maximum likelihood estimation. The key to such an approach is that the information containing which event triggers which subsequent event is unknown and is therefore treated as probabilistic. The expectation step (E-step) updates the probability matrix consisting of the overall branching structure based on $\hat{\theta}^n_{EM}$. The maximization step (M-step) determines the updated parameter estimates, i.e. $\hat{\theta}^{n+1}_{EM}$ based on the updated branching structure computed in the E-step. The EM algorithm is run until convergence (Veen and Schoenberg, 2008). The probabilities of such triggerings, for each pair of points, are updated iteratively as parameter estimates are also updated (Veen and Schoenberg, 2008).

In 2008, Marsan and Lengliné extended this EM-based method to include non-parametric optimization when analyzing seismic activity. They assumed that the triggering function was stepwise, which allowed for estimation of the step heights as though they were parameters in a parametric model. This was in essence estimating the step heights using the EM algorithm, but also allowed for the added flexibility of a data-driven process.

Since the initial use of the EM algorithm proposed in Marsan and Lengliné (2008) was intended for seismic applications, $w_{i,j}$ is defined to be the weighted influence that earthquake i had in triggering earthquake j. The entire matrix of weights is lower triangular in nature where the entry for row i and column j is $w_{i,j}$ and the diagonal entries represent the probability that earthquake i was a main event. Assuming this definition, the expectation step of the EM algorithm involves computing the updated matrix of weights given the spatial-temporal conditional intensity $\lambda(x,t)$. It is designated in Marsan and Lengliné (2008) as

$$w_{ij} = \alpha_j \lambda(|x_j - x_i|, t - t_i, m_i) \tag{2.2}$$

where α is a normalizing coefficient such that $\sum_{i=1}^{j-1} w_{i,j} + w_{0,j} = 1$.

The weights are updated based on the conditional rate and density of earthquakes of a specified magnitude at a given time and location. They are also normalized such that the sum of the following two components add up to one; A) the weights (probabilities) of the $j-1$ earthquakes influencing earthquake j and B) earthquake j being a main event (Marsan and Lengliné, 2008).

In this setting, the expectation step of the EM algorithm involves computing an updated matrix of triggering probabilities given the conditional intensity. In the maximization step, estimates of λ are recalculated given the updated matrix of triggering probabilities (Marsan and Lengliné, 2008). The EM algorithm is iterative and runs until convergence. The orig-

inal application was in estimating seismic activity in the greater Los Angeles area during the 1984-2002 period where the algorithm converged quickly and resulted in an accurate model taking into account a previously studied pattern of cascading aftershocks (Marsan and Lengliné, 2008).

While the initial use for the algorithm in Marsan and Lengliné (2008) was for earthquake analysis, more recent studies have applied it when analyzing purely temporal epidemiologic processes. For instance, Park et al. (2022) applied a purely temporal form of the EM algorithm introduced in Marsan and Lengliné (2008) to fit a simpler non-parametric Hawkes model describing the spread of Ebola during the 2014 epidemic in Africa. It was found that the non-parametric fit of the Hawkes model delineated the spread of the virus as well or better than the traditional SEIR models used for the Ebola outbreak (Park et al., 2022). However, the non-parametric EM algorithm used in Park et al. (2022) was only used to fit the traditional Hawkes model. A potential weakness in using the basic Hawkes model is that any continuation of the disease relies on the assumption of constant productivity throughout the course of the outbreak (Park et al., 2022). There may be a decrease in further disease outbreaks since each occurrence of the disease only directly triggers a Poisson(κ) number of new outbreaks where $\kappa < 1$ (Park et al., 2022). Thus, such a model might underestimate the disease spread early on when the cumulative number of subjects infected is low and overestimate it during the later stages of the epidemic when the disease has already affected a population (Park et al., 2022). Hence, we build on this body of research by introducing a non-parametric form of the recursive model (Schoenberg et al., 2019) taking variable productivity into account.

2.3 Methods

2.3.1 Non-Parametric Estimation of the Recursive Model

In fitting the recursive model (2.1), both the triggering function g and the productivity function H must be estimated. We propose an iterative EM-based procedure similar to that in Marsan and Lengliné (2008), where the productivity function H is estimated based on initial estimates of the background rate μ and the triggering function g. This estimated productivity function is then used to update estimates of the background rate μ and the triggering function g, and so on until a level of convergence is reached.

A few details should be given before describing the algorithm. We suppose purely temporal point process data of the form $\tau_1, \tau_2, ..., \tau_n$, where n is the number of events observed in the time interval $[0, T]$. P is defined as a matrix of estimated probabilities such that P_{ij} is the estimated probability that event i was directly triggered by event j. This probability matrix is by definition lower-triangular since an event j can only trigger later events. Each diagonal entry P_{ii} represents the probability that infection i is a background event, not triggered by any previous event. The sum of any row of P must therefore equal 1, since each event must either be a background event or have been triggered by some prior event. As in Marsan and Lengliné (2008), g is assumed to be a step function indicating the density of triggered points with time lags in some prespecified bins, and similarly we estimate H as a step function with predefined bins corresponding to intervals of the conditional rate, λ. Each of the step heights for h is estimated by determining the productivity in each bin where productivity is defined as the sum of the proportion of events that result directly from prior events rather than entirely new main events. The strategy is to estimate the step heights for g and H as though they were parameters as in Marsan and Lengliné (2008) to allow for both the increased flexibility of non-parametric estimation as well as easier interpretation and forecasting ability.

The number of bins chosen for both g and H is user defined. There seems to be a bias/variance tradeoff between having too many bins or too few, and a possible rule of thumb is to choose the bins so that there are at least 100 time points per bin. In this application, with 13,627 points in our training dataset, we chose 100 bins for g and H. In addition, to ensure that there are sufficient observations in each bin for g and H, all of the smallest interevent times are grouped together into one bin for g and all of the highest productivities are similarly grouped together into one bin for H. We assume here that the background rate μ is constant and estimate this parameter μ as well.

After a quick initialization step (first E-step) where initial estimates of P_{ij} and H are defined, the following EM algorithm can be run until a level of convergence or number of iterations is reached. Note that the initial value for H is the constant $\sum_i \sum_{j<i} \frac{\hat{P}_{ij}}{n}$, which is the sum of the probabilities that each infection is triggered by some preceding point rather than a new immigrant event, which is a sensible initial estimate for the productivity.

Maximization Step:

Part 1: Estimate $\hat{g}$. First determine the interevent times, $\tau_i - \tau_j$, for all positive integers $j < i < n$. For each bin B_l, an interval of the real line containing some of the interevent times, set

$$\hat{g}_l = \frac{\sum_i \sum_{j<i} I(\tau_i - \tau_j \in B_l)(\hat{P}_{ij})}{\sum_i \sum_{j<i} w_l \hat{P}_{ij}} \tag{2.3}$$

where w_l is the width of bin B_l. The bins B_l need not be of equal width. We suggest setting the bins so that they span the entire range of interevent times. To be consistent with the methodology used for the non-parametric version of the Hawkes model (Gordon, 2017), a loglinear approach using base 10 is applied here.

Part 2: Update $\hat{\mu}$.

$$\hat{\mu} = \frac{1}{T} \sum_{i}^{n} \hat{P}_{ii}. \tag{2.4}$$

Part 3: Update $\hat{\lambda}(\tau_i)$ using $\hat{\mu}$ and the most recently updated estimates of $\hat{g}$ and $\hat{H}$, letting

$$\hat{\lambda}(\tau_i) = \hat{\mu} + \sum_{j=1}^{i} \hat{H}_j \hat{g}(\tau_i - \tau_j). \tag{2.5}$$

Part 4: Update estimates of the productivities $H_i = H(\lambda(\tau_i))$. For each bin C_k, an interval of the real line containing some of the values of $\hat{\lambda}(\tau_j)$, set

$$\hat{H}_k = \frac{\sum_{j=1}^{n-1} \sum_{i=j+1}^{n} I(\hat{\lambda}_j \in C_k)(\hat{P}_{ij})}{\sum_{j=1}^{n} I(\hat{\lambda}_j \in C_k)}, \tag{2.6}$$

provided $\sum_{j=1}^{n} I(\hat{\lambda}_j \in C_k) > 0$, and $\hat{H}_k = 0$ otherwise. Thus the sum of the columns of $\hat{P}_{ij}$ where $\hat{\lambda}_j \in C_k$ are averaged. Then, for any j such that $\hat{\lambda}(\tau_j) \in C_k$, set $\hat{H}(\lambda(\tau_j)) = \hat{H}_k$.

Expectation Step:

Update $\hat{P}$ using the values of $\hat{g}$, $\hat{H}$ and $\hat{\mu}$ obtained in the Maximization Step.

$$\hat{P}_{ij} = \frac{\hat{g}(\tau_i - \tau_j)\hat{H}_j}{\hat{\mu} + \sum_{k=1}^{i-1} \hat{g}(\tau_i - \tau_k)\hat{H}_k}. \tag{2.7}$$

Running all the parts of the expectation and maximization steps results in one complete iteration of the recursive Algorithm. While it may be customary to run until convergence, due to small changes from iteration to iteration in the large matrix P_{ij} of size $n \times n$, the algorithm often fails to converge completely. In such cases, we terminated the algorithm

after 100 iterations.

2.3.2 Fitting and Forecasting Point Process Models

When fitting a point process model, the exact inter-event times between events are of importnace in determining the triggering function. Since the dataset of interest consists of weekly state totals, the onset time for each infection within a given week period is randomly drawn from a uniform distribution covering the 7 day time interval as in (Park et al., 2022) and (Schoenberg et al., 2019). In other words, the randomly assigned onset time within a given week is considered to be the exact event time for the purposes of this analysis.

Once infection times are randomized, the interevent times are then used to estimate the triggering functions to fit each model. In the case of the parametric versions of the point process models, the values for the parameters in the triggering functions are estimated via maximum likelihood estimation (Ogata, 1978). For the parametric Hawkes process,

$$\lambda(t) = \mu + \kappa \sum_{i:\tau_i<t} g(t - \tau_i) \tag{2.8}$$

two common forms for the triggering function g are the Pareto

$$g(u) = (p-1)c^{p-1}\frac{1}{(u+c)^p} \tag{2.9}$$

and the exponential

$$g(u) = \beta e^{-\beta u}. \tag{2.10}$$

When fitting the parametric recursive model, the productivity κ varies by

$$\kappa_i = \frac{c}{(\lambda_i)^p} \tag{2.11}$$

To fit the non-parametric Hawkes model, we used the version of the Marsan and Lengliné (2008) EM algorithm implemented by Gordon (2017), which is available in the *R* package *nphawkes*.

Once the point process models are fit to the data, the next step is to forecast future events using the models obtained from the training dataset. In this analysis, we project the number of cases one week into the future using a fitted model of choice as well as actual case counts from weeks prior. Here, we forecast using a simulation method first introduced by Lewis and Shelder (1979). The thinning method works by first generating a homogeneous Poisson process of candidate points with rate b, a large number which should be greater than the maximum conditional intensity observed in the training data. Then, after sorting all the candidate points $(\tau_i, i \in 1, ...b)$ in chronological order, we keep each of these points independently from all others with probability $\frac{\lambda(\tau_i)}{b}$. The conditional intensity for the forecast period of interest would then be estimated using the sorted collection of points consisting of all event times in the training data, kept points from any prior forecast wekks and accepted candidate points from the current period using a fitted point process model of choice with the parameters or bins kept static. This method works for all point process methods considered.

To assess and compare competing models, one may inspect the likelihood as well as the root mean squared error (RMSE) over each week. The latter is particularly relevant when making comparisons over a testing period for simulations of the models fit using separate training data. Because of occasional outliers in the simulations, we compared the trimmed mean of the simulated totals, with the top and bottom 10% of these values removed, to the observed weekly totals in the testing data. Another statistic useful for goodness of fit assessment is the scaled Stoyan-Grabarnik statistic (Stoyan and Grabarnik, 1991), $\sum_{i=1}^{n} \frac{1}{T\hat{\lambda}(\tau_i)}$, which should ideally be close to one if the estimates $\lambda(\tau_i)$ are close to the true conditonal rate, since the expected value of the statistic is easily seen to equal one by the martingale

formula (Baddeley et al., 2005).

Another diagnostic tool useful for comparing the fit of point process models is superthinning (Clements et al., 2012). Superthinning is a method of combining thinning observed points (Lewis and Shelder, 1979) in areas with high intensity and superposition of simulated points (Brémaud, 1981) in sparse regions to perform a diagnostic check on the residuals. According to Clements et al. (2012), superthinning results in a homogeneous residual point process only if the conditional intensity of the fitted model estimates the actual values precisely. To avoid the possible pitfalls of thinning or superimposing, a combined method is used for superthinning (Clements et al., 2012). Given a constant b selected by the user, such as the mean of the estimated conditional intensities at the observed points (Gordon et al., 2015), one thins the data keeping each observed point τ_i independently with probability $\min\{1, \frac{b}{\hat{\lambda}(\tau_i)}\}$ and then superposing simulated points according to a Poisson process with rate $\max\{0, b - \hat{\lambda}(t)\}$. If the estimated conditional rate $\hat{\lambda}(t)$ is correct, then the resulting residual process is a stationary Poisson process of rate b. The superthinned residuals may thus be inspected for trend, clusters, gaps or other patterns as evidence of lack of fit of the model.

2.4 The SVEILR Model

The method to fit and forecast the SEIR model differs substantially from that of the point process model. While no special considerations have to be made for point process models regarding a specific disease, details such as the rate of exposure, method of spread and human intervention must be taken into account in the design of a compartmental model. This is because the model is designed as a closed system of differential equations where subjects go from one state, i.e. susceptible, infected or recovered to another at a rate determined by terms that have a simple interpretation (Kermack and McKendrick, 1927). A single vaccine

model is warranted since the original mumps vaccine was introduced before this period in 1967 and the second dose was not mandated until 1989 (Centers for Disease Control and Prevention, 2019).

We followed Li et al. (2018), who developed a one vaccine model for mumps in mainland China, where the disease is still prevalent. Although there is sufficient availability of the second MMR dose, China only provides one free dose of MMR and there is no push from National Health and Family Planning Commission of the People's Republic of China (NHFPC) to require children to take a second dose (Li et al., 2018). The proposed expanded SVEILR (Susceptible, Vaccinated, Exposed, Infected, Light Infection, Recovered) model adds two necessary nodes to take into account the vaccinated state (V) as well as a lightly infected state (L) since not all cases are symptomatic (Li et al., 2018), (Centers for Disease Control and Prevention, 2019). However, with waning immunity, a subject can either return to the susceptible state (S) or even be exposed (E) to the virus (Li et al., 2018).

In this analysis, the differential equation model used to model mumps in Pennsylvania during the 1970-1990 time period is quite similar to the one applied in Li et al. (2018). As in (Li et al., 2018), the values optimized by minimizing the sum of squares include the transmission rate (β), waning immunity rate (λ), vaccine coverage of the susceptible/exposed (ϵ), proportion of people seeking medical advice (γ) and initial proportions susceptible and vaccinated ($S(0), V(0)$).

The system of differential equations developed by (Li et al., 2018) for this SVEILR model is

as follows:

$$\begin{aligned}
\frac{\partial S}{\partial t} &= \mu - p\mu E - \beta S(I+L) + \lambda V - (\epsilon + \mu)S \\
\frac{\partial V}{\partial t} &= \epsilon S + \epsilon_1 E - \lambda V - \kappa\beta V(I+L) - \mu V \\
\frac{\partial E}{\partial t} &= \beta S(I+L) + \rho\mu E + \kappa\beta V(I+L) - (\alpha + \epsilon_1 + \mu)E \\
\frac{\partial I}{\partial t} &= \alpha\gamma E - (\delta_1 + \mu)I \\
\frac{\partial L}{\partial t} &= \alpha(1-\gamma)E - (\delta_2 + \mu)L \\
\frac{\partial R}{\partial t} &= \delta_1 I + \delta_2 L - \mu R
\end{aligned}$$

The set of differential equations establish the rate that the numbers in each state change over time. In addition, each term in this model represents an effect which can be interpreted independently. As described in Li et al. (2018),

$$\begin{aligned}
\beta &= \text{transmission rate} \\
\lambda &= \text{waning immunity rate} \\
\epsilon &= \text{vaccine coverage of the susceptible} \\
\epsilon_1 &= \text{vaccine coverage of the exposed} \\
\kappa &= \text{invalid vaccination rate} \\
\alpha &= \text{rate moving from exposed to severe or mild infectious} \\
\gamma &= \text{proportion of the severe infections seeking medical advice} \\
\delta_1 &= \text{rate moving from severe infectious to recovered} \\
\delta_2 &= \text{rate moving from light infectious to recovered.}
\end{aligned}$$

The model was then applied to monthly mumps data acquired from the CDC of China. Since mumps spreads the most rapidly amongst children in school, Li et al. (2018) fit the SVEILR model for half year periods since peak cases tend to occur after major holiday breaks. The

SVEILR model is fit in Li et al. (2018) using data from February 2009 to September 2014 by minimizing the sum of squares between the number of severe infections (state I) accounted for by the model and the actual case count and then forecast from October 2014 to September 2015.

The only difference in our fitting of the model to mumps in Pennsylvania versus that of Li et al. (2018) is that peaks in mumps cases occur in Pennsylvania annually corresponding with the beginning of the academic year when schoolchildren return from break rather than twice per year as reported in (Li et al., 2018). As a result, the SVEILR model is fit here using weekly data for each year separately beginning from the first week in October and ending in the last week in September.

Forecasting the SVEILR model was performed using standard exponential smoothing. Parameter estimates are obtained by taking a weighted average for all the years during the training period (Shmueli and Lichtendahl, 2016). To smooth, a reasonable value of the smoothing parameter α is chosen and for each of the six parameters of interest $(\beta, \lambda, \gamma, \epsilon, S(0), V(0))$, the value of each parameter estimate is determined by $(\frac{1-\alpha}{\alpha})(\alpha p_{i-1})(\alpha^2 p_{i-2})...(\alpha^k p_{i-k})$ where p_i is the parameter estimate from year i.

2.5 Data

Recorded cases of mumps statewide in Pennsylvania between January 1970 and December 2001 were obtained courtesy of Project Tycho (Van Panhuis et al., 2018). The dataset consists of weekly statewide case totals during this time period. According to Centers for Disease Control and Prevention (2019), mumps is a contageous viral disease, usually spread through airborne transmission. Initial symptoms of mumps include fever, headache, fatigue and loss of appetite and then subsequent swelling of the salivary glands (Centers for Disease

Control and Prevention, 2019). While morbidity from mumps is low (2 out of 10,000 cases in the United States in the pre-vaccine era), mumps has been known to cause orchitis (testicular inflammation), encephalitis and even permanent hearing loss in a small percentage of cases (Centers for Disease Control and Prevention, 2019).

In 1967,the still currently used mumps vaccine was introduced in the United States. In the U.S., the mumps vaccine is now available in combination with the measles and rubella vaccines, known as MMR (Centers for Disease Control and Prevention, 2019). In 1977, one dose of the MMR vaccine was recommended for all children 12 months and older. However, in 1989, the vaccine policy was updated to include a second dose at 4-6 years of age after it was discovered that waning immunity to mumps was a real possibility (Centers for Disease Control and Prevention, 2019). As a result, the incidence of mumps in the U.S. has dropped from 55.5 cases per 100,000 people to < 2 cases per 100,000 in 2017 (Elflein, 2019).

Despite the fact that the MMR vaccine has done wonders to slow down a once all too common disease, it is important to continue studying the prevalence of mumps since occasional outbreaks still occur. For example, in 1986 and 1987, there was a resurgence of mumps, primarily among teenagers born prior to the 1977 MMR vaccine mandate (Centers for Disease Control and Prevention, 2019). Additionally, although the MMR vaccine has been available for decades, sporadic outbreaks of mumps still happen. In 2006, there was an outbreak which affected many midwestern university students, most of whom lived in dormitories (Centers for Disease Control and Prevention, 2019). Another occurred in 2009 when 3502 cases of mumps were reported in the Orthodox Jewish communities in New York City, an area where 90% of the children at the time received one or more doses of MMR (Centers for Disease Control and Prevention, 2019). These mini-resurgences of mumps have led researchers to question whether those vaccinated are still prone to losing immunity to the disease over their lifetimes.

The dataset of interest includes weekly Pennsylvania state totals for reported cases of mumps over a 32 year period beginning in January 1970. Any weeks with no data over this time-frame are treated as having no confirmed cases. Figure 2.1 illustrates the weekly incidence of mumps statewide for the entire time period.

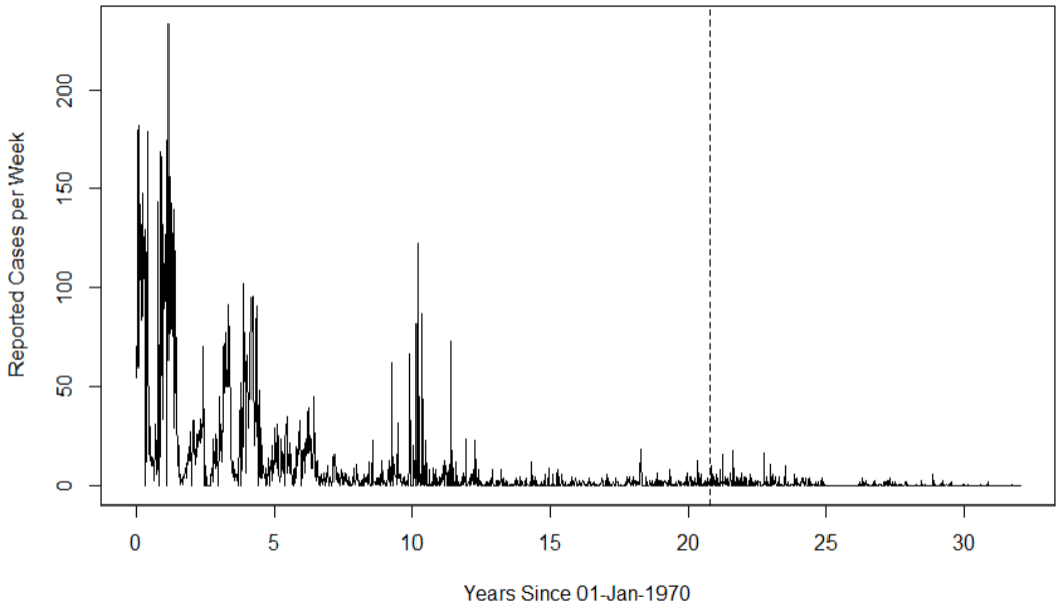


Figure 2.1: Weekly reported cases of mumps in Pennsylvania over time. To the left of the dotted line is the training data which includes all case counts from January 1970 to September 1990. To the right is the testing data which includes all weekly counts from October 1990 to December 2001.

In order to evaluate the fitting and forecasting ability of each of the models, we use the data recorded for the first $\frac{2}{3}$ of the time frame of interest, which includes all weekly case counts from January 1970 to September 1990, as the training portion. The testing portion is the last third of the dataset, or weekly reported infections from October 1990 to December

2001.

2.6 Results

2.6.1 Model Fitting

Table 2.1 lists the parameter estimates for the SVEILR model for each of the twenty years fit individually during the course of the training period.

Academic Year	β	λ	γ	ϵ	S(0)	V(0)	I/100000
Jan 1970 - Sep 1970	0.55211	0.01426	0.16893	0.00100	0.08833	0.61189	68
Oct 1970 - Sep 1971	0.34540	0.01142	0.19802	0.00100	0.07234	0.75687	51
Oct 1971 - Sep 1972	0.56523	0.00521	0.05487	0.00100	0.02342	0.82246	40
Oct 1972 - Sep 1973	0.48028	0.00615	0.08134	0.00100	0.02063	0.87002	49
Oct 1973 - Sep 1974	0.38864	0.00915	0.10813	0.00100	0.07885	0.75271	41
Oct 1974 - Sep 1975	0.32220	0.00858	0.04812	0.00100	0.04734	0.72389	36
Oct 1975 - Sep 1976	0.87266	0.00618	0.07130	0.00100	0.00391	0.54681	35
Oct 1976 - Sep 1977	0.58590	0.00502	0.01516	0.00100	0.06679	0.82386	12
Oct 1977 - Sep 1978	0.52773	0.00452	0.03159	0.00621	0.00001	0.78110	33
Oct 1978 - Sep 1979	0.33159	0.00704	0.02371	0.00100	0.07724	0.86552	16
Oct 1979 - Sep 1980	0.41620	0.01561	0.05648	0.00943	0.00001	0.93898	17
Oct 1980 - Sep 1981	0.42398	0.02703	0.05925	0.03364	0.00001	0.98814	63
Oct 1981 - Sep 1982	0.48071	0.02221	0.05562	0.04158	0.00001	0.95661	13
Oct 1982 - Sep 1983	0.38916	0.02290	0.01558	0.02730	0.00001	0.92277	5
Oct 1983 - Sep 1984	0.31616	0.00602	0.00599	0.00100	0.20079	0.74775	4
Oct 1984 - Sep 1985	0.38391	0.01970	0.05065	0.03394	0.08606	0.89321	8
Oct 1985 - Sep 1986	0.69370	0.06991	0.13761	0.11711	0.88178	0.11712	5
Oct 1986 - Sep 1987	0.25167	0.02881	0.02315	0.02672	0.00001	0.95429	4
Oct 1987 - Sep 1988	0.48329	0.00521	0.01040	0.00100	0.06763	0.83424	8
Oct 1988 - Sep 1989	0.12743	0.02214	0.09552	0.00790	0.15226	0.82994	8
Oct 1989 - Sep 1990	0.79982	0.00534	0.01344	0.00100	0.00001	0.58238	8

Table 2.1: Annual estimates of the transmission rate (β), waning immunity rate (λ), vaccine coverage of those who are susceptible or exposed (ϵ), proportion of people with severe infections (γ), initial proportions susceptible and vaccinated ($S(0), V(0)$) and rate of severe infection per 100,000 people at the end of each year ($\frac{I}{100000}$).

The decrease over time in the annual estimated incidence of infection is obvious. This may be due to corresponding increases in the vaccination rate which rose from 58.4% in 1970 to a high of 67.6% in 1982 (Centers for Disease Control and Prevention, 2011). However, data from 1986 to 1990 was not recorded and the survey methods to collect such data changed after 1993 resulting in dramatically higher vaccination rates being reported (Centers for Disease Control and Prevention, 2011). Note that some of the reduction in reported cases is also reflected by a drop in the estimates of the proportion of infections that are severe, γ. This is especially apparent in the precipitous drop in $\hat{\gamma}$ from 1970 to 1971.

In addition, Figure 2.2 shows the fit of the SVEILR model to the Pennsylvania mumps training data on the left side of the dotted line.

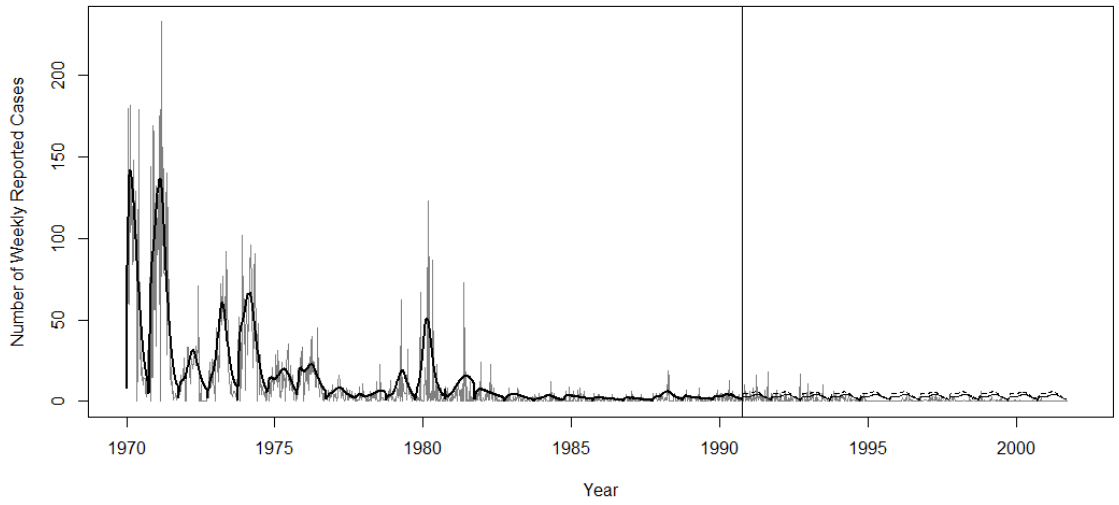


Figure 2.2: Reported weekly total cases of Pennsylvania mumps (gray) with fitted SVEILR model estimates for the training period (solid black) and SVEILR forecasts for the testing period using exponential smoothing with $\alpha = 0.05$ (dash line) or $\alpha = 0$ (solid line).

The SVEILR model appears to adequately account for both the seasonality of mumps infections as well as the decreasing overall trend during the twenty year period. In addition, the model is largely identifying the peak times correctly. The RMSE for the SVEILR model over the training period is 15.58 cases per week.

For the simple Hawkes model given by (2.8) and with triggering function (2.9), the following parameter estimates are obtained using MLE: $\hat{\mu} = 0.089$, $\hat{\kappa} = 0.951$ points per observed event, $\hat{c} = 1.82$ events per day, and $\hat{p} = 2.50$. For the recursive model with triggering function (2.10) and with recursive component (2.11), the estimates are $\hat{\mu} = 0.1279$, $\hat{c} = 0.8623$ points per observed event, $\hat{\beta} = 0.7027$ events per day, and $\hat{p} = -0.04486$.

Fitted estimates of the triggering function, g, for the parametric and non-parametric Hawkes and recursive models are shown in the top panel of Figure 2.3. The bottom panel shows the corresponding estimated productivity functions, $\hat{H}(\lambda)$, for the four models.

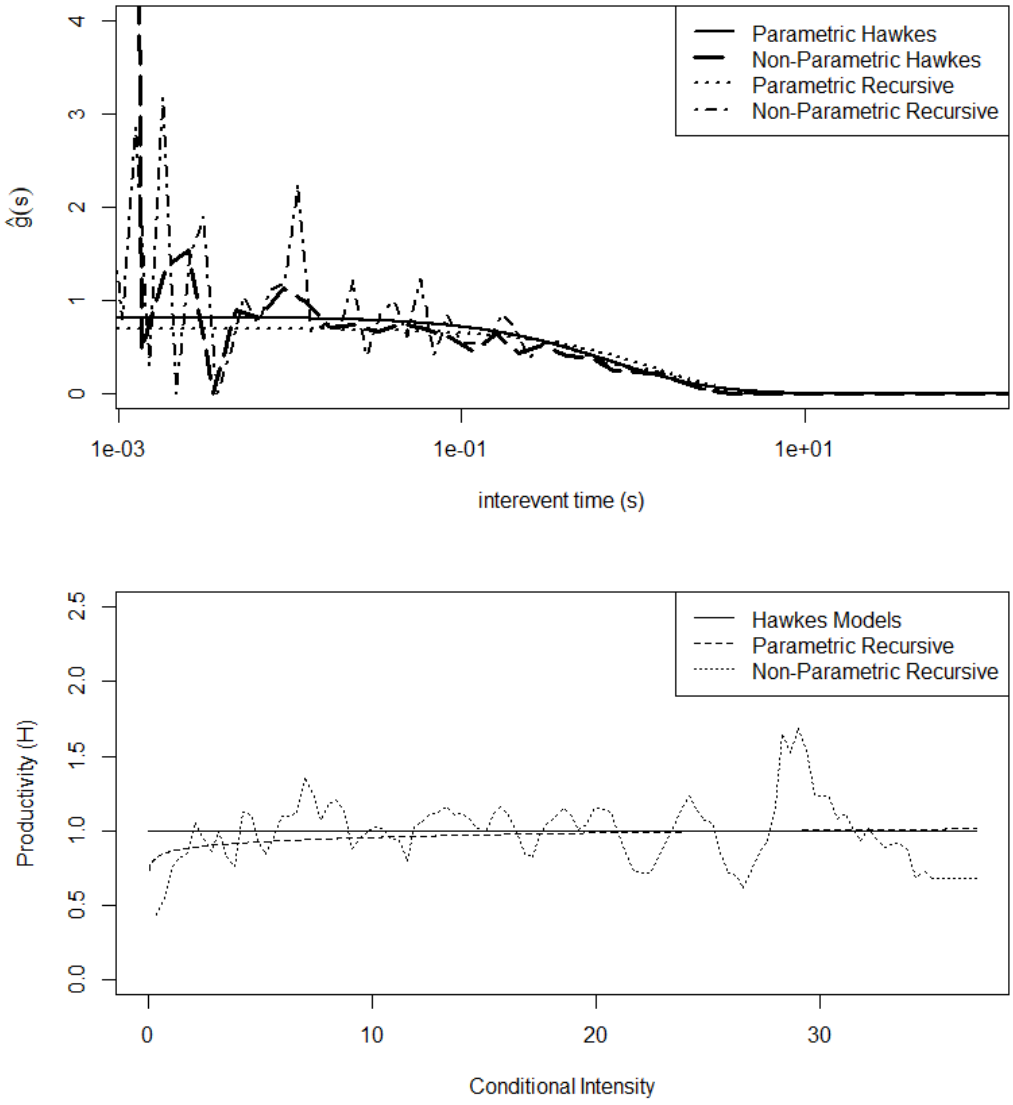


Figure 2.3: Estimated triggering density, g, for the parametric Hawkes, non-parametric Hawkes, parametric recursive and non-parametric recursive models (top). Estimated productivity function, $\hat{H}(\lambda)$, for the Hawkes, parametric recursive and non-parametric recursive models (bottom).

There appear to be rather substantial variations in productivity as the estimated conditional rate varies.

Figure 2.4 shows the weekly numbers of observed cases on mumps in Pennsylvania along with the fitted rates from the fitted parametric and non-parametric Hawkes and recursive models. All four of the point process models appear to fit quite closely on the training data, despite underestimating the largest peaks, especially in 1970-1972.

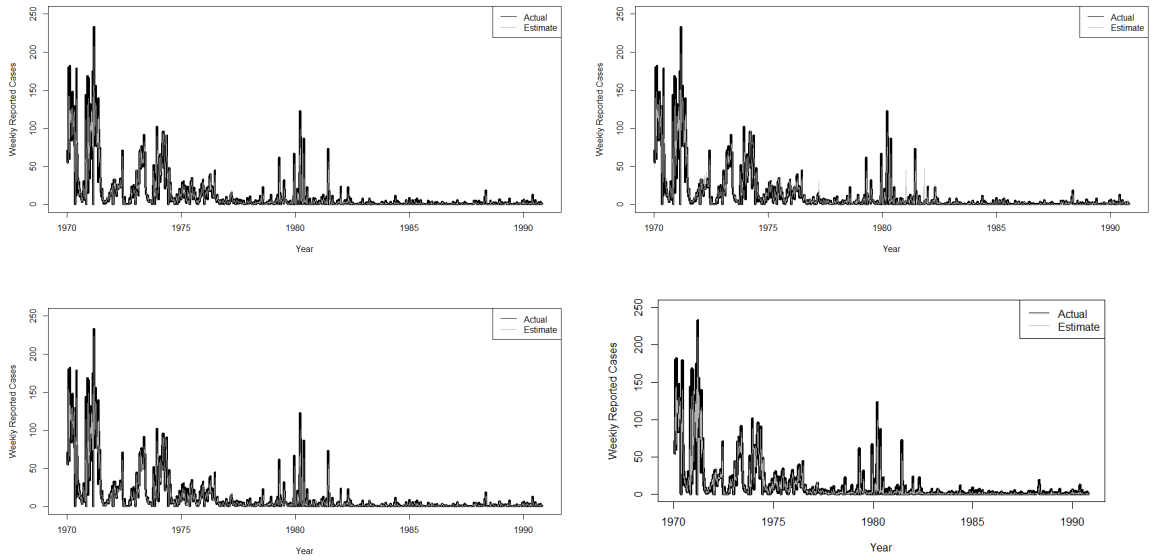


Figure 2.4: Estimated conditional rates and observed weekly number of cases of mumps in Pennsylvania over the 1970-1990 training period. The fits for the parametric Hawkes (top left), non-parametric Hawkes (top right), parametric recursive (bottom left) and non–parametric recursive (bottom right) models are shown along with the observed number of reported cases.

As shown in Table 2.2, the non-parametric version of the recursive model achieves the lowest

RMSE of five models considered.

Model	Training RMSE	Stoyan-Grabarnik	Forecasting RMSE	Standard Error
SVEILR ($\alpha = 0.05$)	15.580	—	4.136	0.095
SVEILR ($\alpha = 0$)	15.580	—	3.059	0.087
Param. Hawkes	5.901	0.9992	2.380	0.090
N.P. Hawkes	5.058	0.9988	2.253	0.089
Param. Recursive	4.744	1.0035	2.502	0.089
N.P. Recursive	4.732	1.0002	1.866	0.077

Table 2.2: RMSE for each model using the training data and the testing data. The standard error reported is for the forecasting RMSE.

All four of the point process models fit the training data substantially better than the SVEILR model. The Stoyan-Grabarnik statistics, for the parametric Hawkes model, the nonparametric Hawkes model, the parametric recursive model, and the nonparametric recursive model, respectively, were 0.9992, 0.9988, 1.0035 and 1.0002, indicating no noticeable lack of fit for any of the four point process models.

2.6.2 Residual Analysis

Figure 2.5 shows the superthinned residuals over time for the non-parametric recursive model. The number of superthinned residuals from 1970 to 1975 is slightly lower than for subsequent years, indicating overestimation of the rate during this early period. Figure 2.5 also shows a histogram and lag plot of the standardized interevent times for the superthinned residuals from the non-parametric recursive model.

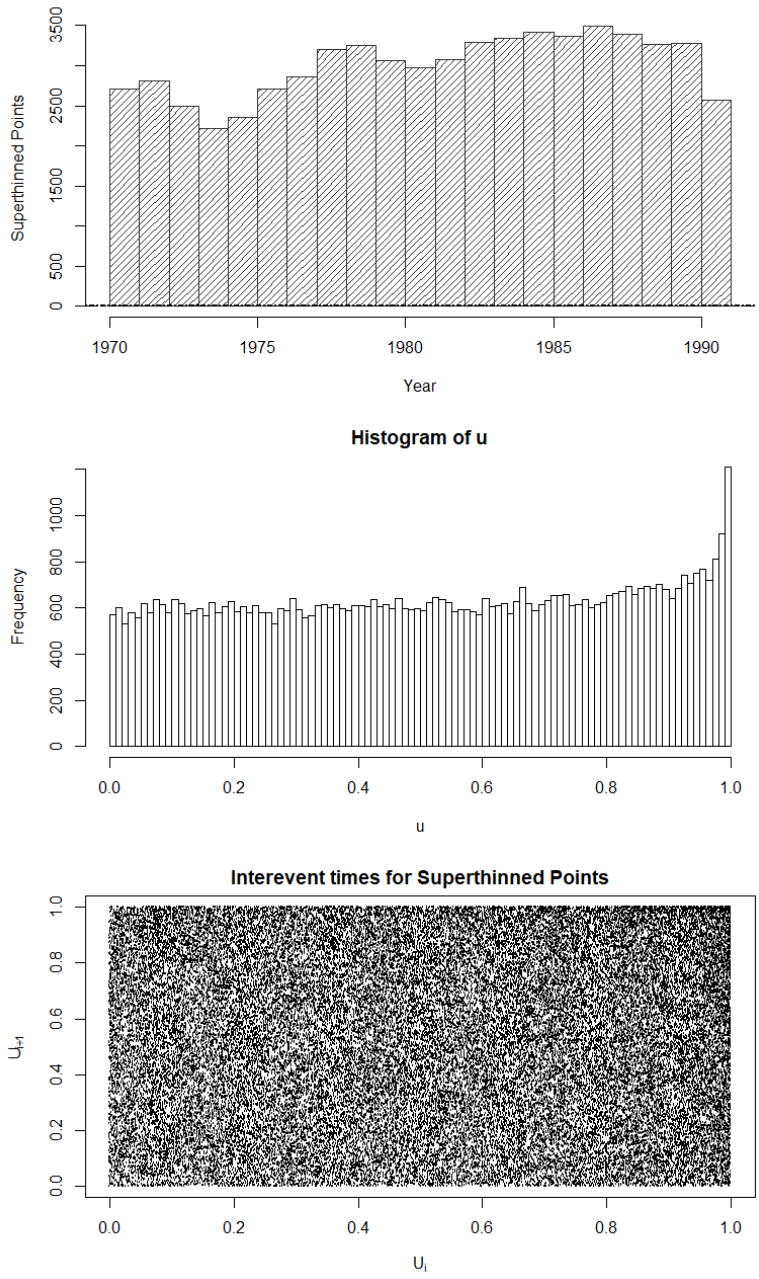


Figure 2.5: For the non-parametric recursive model: (a) Count of superthinned residuals over time using $b = 100$, (b) histogram of standardized times between consecutive events, $u_i = F^{-1}(\tau_i - \tau_{i-1})$, where F is the cumulative distribution function of the exponential with rate $b = 100$, and (c) lag plot of standardized times between consecutive events, for superthinned residuals.

There is a noticeably higher frequency than expected of the longest interevent times for the non-parametric recursive model. However, there is no noticeable clustering of the interevent times.

2.6.3 Out-of-Sample Forecasting

Figure 2.2 shows forecasts using the SVEILR model with parameters exponentially smoothed using $\alpha = 0.05$ or with no smoothing ($\alpha = 0$). The weekly reported case counts during the testing period resemble the last few years in the training period more than the earlier years. To note, the RMSE is lower using $\alpha = 0$ than when using $\alpha = 0.05$.

The parametric Hawkes, parametric recursive and non-parametric Hawkes models forecast the number of cases accurately from 1990-1994, but substantially overestimate the cumulative number of infections during later years. On the other hand, the forecast using the non-parametric recursive model slightly underestimates the case counts during the first several years of the testing data but is relatively accurate from 1996-2001. Overall, as shown in Table 2.2, the non-parametric recursive model has the smallest RMSE of 1.866 cases/week, followed by the non-parametric Hawkes model with an RMSE of 2.253 cases/week. The SVEILR model, whether with $\alpha = 0$ or $\alpha = 0.05$, had substantially higher RMS errors.

The non-parametric recursive model has both the lowest RMSE within the training portion of the dataset as well as improved accuracy in forecasting weekly cases of mumps in the testing dataset. The next best in fitting was the parametric recursive and the runner up for forecasting was the non-parametric Hawkes model. The RMS error for all four point process models was smaller than that of the SVEILR model for both fitting and forecasting.

2.7 Discussion

In this chapter, we have introduced a non-parametric version of the novel recursive point process model first introduced in Schoenberg et al. (2019) for epidemiologic application. We show that the nonparametric recursive model appears to outperform the alternative models for forecasting weekly mumps cases in Pennsylvania and is also substantially less prone to overfitting as evidenced by the improved performance during the testing period relative to competing point process and SVEILR models.

An important item for further research would be to explore how all of these models might be improved by estimating a nonstationary background rate, incorporating the decreasing trend in cases as well as their seasonality, while simultaneously estimating the triggering portions of the models as in Zhuang and Mateu (2019), for example. Additionally, future work should focus on ideal selection of bin widths and bin numbers for fitting the nonparametric forms of the recursive and Hawkes models, especially for the estimation of the productivity function, H. Also of importance is the search for optimal ways to smooth the simple binned estimates considered here, perhaps via kernel smoothing as in e.g. Mohler (2014). Furthermore, future research should include adapting both the non-parametric as well as the parametric comparison models to take into account the more recent HawkesN model advance as in (Rizoiu et al., 2018), (Kong et al., 2020) and (Kresin et al., 2021), which allows for a finite susceptible population.

The methods explored here should be applied to other diseases as well, including emerging epidemics. It is unclear as to whether the techniques of fitting and forecasting would remain the same if the incidence of a disease is static or even increases over time.

Another important potential use of the non-parametric recursive model is to forecast the prevalence of various diseases based on the rate of vaccine coverage. This is particularly rel-

evant in light of recent anti-vaccination trends in many developed nations, some of which may be attributable to misguided fears about unknown side-effects from vaccines (Dubé et al., 2015). According to (Dubé et al., 2015), 5% of parents in the United States refuse to have their children vaccinated and up to one-third of children in the U.S. lack the recommended protection from easily avoidable diseases.

CHAPTER 3

Doubling Time Estimation Using the HawkesN Model for SARS-COV-2 in California

3.1 Introduction

Epidemiologists frequently refer to the doubling time of an infectious disease as a way to summarize how rapidly the disease is spreading in a particular location. Doubling time is defined as the time needed for total infections to double (Centers for Disease Control and Prevention, 2009), (Centers for Disease Control and Prevention, 2021a) and is easily interpretable by both epidemiologists and the general public. It is also a useful statistic relating how the number of cases is expected to change as a function of time during a surge of cases. Despite this importance, some ambiguity exists in the definition of doubling time. Furthermore, the relationship between doubling time and parameters in widely used compartmental and point process epidemic models is not fully understood. In Merler et al. (2013), the relationship between doubling time and the reproduction number parameter in the Susceptible-Infected-Recovered (SIR) model, R_0, is explored. In this chapter, we extend the results of Merler et al. (2013) by exploring the relationship between doubling time and parameters in the HawkesN point process model (Rizoiu et al., 2018).

As described in Chapter 1, the simple Hawkes model is not a stable (stationary) process whenever there is an increase in cases, or $\kappa > 1$ (Park et al., 2022). However, the HawkesN adaptation (Rizoiu et al., 2018) solves the stability issue for expanding epidemics. The

HawkesN model as shown in Rizoiu et al. (2018)

$$\lambda(t) = (1 - \frac{N_t}{N})[\mu + \sum_{t_j<t} \phi(t - t_j)] \tag{3.1}$$

where $\phi(t - t_j)$ is defined using the exponential kernel function

$$\phi(t - t_j) = \kappa\theta e^{-\theta(t-t_j)} \tag{3.2}$$

assumes a finite susceptible population N and removes any subject who has been previously infected prior to time t. This allows for the conditional intensity at time t to be proportional to the size of the remaining susceptible group at time t, which allows for stability when $\kappa > 1$ (Kresin et al., 2021).

Another unique property of HawkesN is that with an exponential kernel, the expected number of individuals infected in a finite population is close to the estimate from a stochastic SIR model as it approaches stationarity, even for an increasing pandemic when $\kappa > 1$ (Kresin et al., 2021). This allows for the ability to make direct comparisons between various features in HawkesN and SIR compartmental models, such as κ and R_0, which both represent the expected number of infections triggered by a previous event (Kresin et al., 2021).

The motivation for this chapter is to expand upon the ideas presented in Rizoiu et al. (2018) and Kresin et al. (2021) and use the HawkesN model to accurately predict the doubling time for an increasing epidemic. Here, two different doubling time measures are used. Cumulative doubling time is the amount of time for the total number of cases to double (Merler et al., 2013) and daily rate doubling time is the number of days it takes for the daily reported infections to double (Johns Hopkins University and Medicine, 2022). To address our goals, we first run a simulation to determine the effects of the productivity term κ and the generational length coefficient β on both doubling time measures of interest. This allows us to

gain a theoretical understanding as to how each component of HawkesN impacts doubling time. Second, we compare the performance of HawkesN in forecasting doubling times for SARS-COV-2 during three surges in California to that of the SQUIDER (Susceptible, Quarantine, Undetected Infected, Dead, Exposed, Recovered) model (Khan et al., 2020), a simple modification of the basic SIR compartmental model adapted to the original surge of the SARS-COV-2 virus. We evaluate the performance of both models using simulations and empirical doubling times based on daily reported statewise totals for SARS-COV-2 infections in California (California Department of Public Health, 2021b) during three time periods when infections were increasing (February & March 2020, November 2020 and July 2021).

3.2 Methods

3.2.1 Simulation of Doubling Time

3.2.1.1 HawkesN Process Generator

In order to obtain estimates for doubling time, we have developed an algorithm to simulate HawkesN processes given the background rate μ, productivity rate κ, generational length β and susceptible population N. In Kresin et al. (2021), it has been shown that one can simulate a disease process using a SEIR-Hawkes process, taking advantage of both the easy interpretability of terms from the SIR compartmental family of models as well as the point process properties of Hawkes. In Kresin et al. (2021), the conditional intensity of new infections is given by

$$\lambda^E(t) = (1 - \frac{N_t^E}{N}) \sum_{t>t_j^I} R_0 \gamma exp(-\gamma(t - t_j^I)) \tag{3.3}$$

and infection times are generated by

$$P(t_j^I > t_j^E + c) = \int_c^\infty \mu exp(-\mu(s - t_j^E))ds \tag{3.4}$$

In (3.3), the conditional intensity is still a function of the susceptible population, productivity and triggering function as in (3.1), but the SIR parameters representing total infections up to time t ($N^E(t)$), transmission rate (R_0) and infection rate (γ) take the place of the usual HawkesN terms (Kresin et al., 2021). Also, in (3.4), a new infection t_j^I is generated at some time interval c after the previous one dependent on an exponential kernel featuring the rate of exposure μ (different from the HawkesN background rate in (3.1)). It is shown in Kresin et al. (2021) that the SEIR-Hawkes process can be simulated using an iterative process.

In this paper, we develop an algorithm to simulate HawkesN processes using a similar method to the one used in Kresin et al. (2021). The goal is to simulate a HawkesN process until a defined termination time T_{end}. This method simulates a branching process where the first set of accepted points consist merely of background infections randomly scattered from time 0 to T_{end} with rate μ. Then candidate offspring are proposed for each background point and are either accepted or rejected to imitate the triggering function g. The next generation then includes the original background points and the most recent accepted offspring and the process repeats until the branching process from time 0 to T_{end} has been exhausted.

Set $R_0 = \kappa$ and $\gamma = \beta$ for interpretability. After the time of infection has been established for each of the initial background events, each iteration of the branching process is comprised of the following steps, closely resembling the algorithm presented in Kresin et al. (2021):

Part 1: For each accepted point from the previous generation, a, the number of candidate

offspring is determined by drawing a random number $M \sim Poisson(R_0)$. Then for each of M proposed future events, the time of exposure is offset from the ancestor's by $exp(\mu)$. Last, all of the accepted points and candidate points are sorted together in chronological order. These will be known as simulated points, or s. The previously accepted points should also be kept in chronological order independently.

Part 2: For each candidate point, c:

$$\lambda_c = 1 - \frac{N_a}{N} \sum_{i=1}^{N_a} R_0 \gamma exp(-\gamma(t(c) - t(a_i))) \tag{3.5}$$

and

$$\nu_c = \sum_{i=1}^{N_s} R_0 \gamma exp(-\gamma(t(c) - t(s_i))) \tag{3.6}$$

where $t(c)$ is the proposed time of infection for the candidate point, $t(a_i)$ is the event time for accepted point i and $t(s_i)$ is the infection time for simulated point i. Also, N_a is the number of accepted points with an infection time before that of the candidate point and N_s is interpreted similarly, but including all simulated points.

Part 3: Accept or reject each candidate point using Lewis' thinning method (Lewis and Shelder, 1979). Accept the candidate point if

$$D_c \sim Unif[0, 1] < \frac{\lambda_c}{\nu_c}. \tag{3.7}$$

Part 4: To take into account exposure time before infection, the time of infection for each newly accepted point is offset by

$$t(a) = t(a) + exp(\gamma). \tag{3.8}$$

Running all parts of the algorithm results in one generation of the branching process for HawkesN.

3.2.1.2 Doubling Time Using HawkesN Simulator

In order to examine the effects of generation length β and infection rate κ on doubling time, 500 simulations are run for each value of $\kappa \in [1.1, 5]$ and several choices of β using the HawkesN process generator. In this case, we are interested in the median cumulative and daily rate doubling times from the 50th infection recorded, the cutoff used in analyses run by Johns Hopkins University and Medicine (2022). Here, cumulative doubling time is defined as the time it takes to record 50 additional infections. Daily rate doubling time is defined as the number of days until twice as many daily infections happen as on the day when the 50th infection occurs.

3.2.2 Model Fitting

3.2.2.1 The HawkesN Model

In fitting any Hawkes model (1.1) the values for μ, κ and the triggering function g must be estimated. This is conventionally done using maximum likelihood estimation and usually results in estimates with beneficial asymptotic properties including efficiency, consistency and asymptotic normality (Ogata, 1978), (Ogata, 1988). However, when fitting the Hawkes model to reported daily case counts without time of day precision, the least squares technique has been known to work well (Schoenberg, 2021) since it takes advantage of the relationship between Hawkes processes and autoregressive techniques derived in (Kirchner, 2016), (Kirchner, 2017). In Schoenberg (2021), simple Hawkes model parameters are fit to daily case counts covering the first 576 days of the COVID-19 pandemic for all 50 states

using least squares, minimizing

$$\sum_{t=1}^{T}(N(t) - [\mu + \sum_{i=1}^{16} \kappa(t-i)g(i)N(t-i)])^2 \tag{3.9}$$

where T is the total elapsed time of 576 days and $N(t)$ is the number of observed infections on day t. In the non-parametric version, thirty-six 16-day periods are fit, resulting in a value for κ each and the g is a stepwise triggering function with 16 steps (Schoenberg, 2021). The Hawkes model fitted using least squares method is shown to fit SARS-COV-2 data quite well in Schoenberg (2021). It also describes COVID-19 transmission in Indiana using an additional step function (Mohler et al., 2021) and COVID-19 in the United States using additional spatial parameters and mobility data (Chiang et al., 2020) effectively.

In this setting, a form of the non-parametric least squares technique for the HawkesN model is also applied since the dataset of interest involves daily case counts. Since the population of California is $40,129,160$ as of the 2020 census (California Department of Public Health, 2021b) and total reported infections in any COVID surge is likely a small fraction of the total susceptible population, the impact of N is miniscule and thus not added into the least squares minimization. However, instead of obtaining a κ for each 16 day period, our least squares algorithm outputs one value for κ minimizing

$$\sum_{t=1}^{T}(R(t) - [\mu + \kappa \sum_{i=1}^{16} g(i)R(t-i)])^2 \tag{3.10}$$

where $R(t)$ is the number of cases reported on day t. This modification allows for shorter duration training datasets of 1 month or less which is needed when examining doubling time. Also, once the triggering stepwise function g is obtained, a non-linear smoothing function is applied resulting in a singular parameter β. This allows for an estimate of the background rate μ, productivity rate κ and inverse of the generation length β which can be fed in to the doubling time simulator.

3.2.2.2 The SQUIDER Model

Since the epidemiologic application used here involves fitting and forecasting daily reported SARS-COV-2 cases, it is only fair that the SIR model used as a comparison to HawkesN is adapted to fit such data. While many compartmental models have been developed for describing the behavior of the COVID-19 pandemic, the SQUIDER model (Khan et al., 2020) is a relatively simple adaptation to the basic SIR model that has been shown to work well fitting and forecasting the initial SARS-COV-2 surge in the United States in February - May, 2020. Similar to the SIR model, SQUIDER is also designed as a closed system of differential equations where individuals susceptible to the virus travel from one state to another at a rate dictated by terms with relatively simple interpretations.

The expanded SQUIDER (Susceptible, Quarantine, Undetected Infected, Infected, Dead, Exposed, Recovered) model has four additional nodes beyond what is present in the basic SIR model in order to take into account specifics regarding the behavior of the COVID-19 pandemic (Khan et al., 2020). The quarantine state (Q) takes into account subjects who have either been potentially exposed and quarantining at home for the required 10 days (CDC, 2021) or those who are staying at home voluntarily due to stay-at-home orders (Khan et al., 2020). Another factor taken into account by the undetected infected (U) and the undetected recovered / dead (E) states is that not all cases or even outcomes are detected due either to a lack of testing or proper reporting (Bahning et al., 2020), (Khan et al., 2020). Lastly, the (D) state represents the known proportion of individuals who pass away from complications due to COVID-19 (Khan et al., 2020).

In this analysis, the differential equation model used to fit and forecast COVID-19 data in California is similar to the original SQUIDER algorithm presented in Khan et al. (2020). As in Khan et al. (2020), optimization involves minimizing the sum of squares between expected and actual counts for both cumulative infections and cumulative deaths simulta-

neously. The resulting fitted parameters include the rates of transmission (β), testing (δ), recovery (α), known deaths (γ), waning immunity (ρ) and undetected outcomes (ϵ) as well as the original proportion of undetected infecteds, U(0), and a which is the normalized number of new infections in time (Khan et al., 2020). The a term allows for a curvilinear optimization to occur since mixing susceptible and undetected infectious populations at a constant rate does not always make sense in states or cities with small or isolated populations (Khan et al., 2020).

An additional component to the SQUIDER model which is also included in the optimization is the quarantine measure. In Khan et al. (2020), both the time of the initial quarantine as well as the proportion of the population that follow through are included as free parameters. To simplify the model, the proportion that either enters or leaves quarantine is measured by single pulses. For this paper, only the share of the population which has entered or left quarantine during two stay-at-home orders issued in California (Executive Department State of California, 2020), (California Department of Public Health, 2021a) is taken into account. In other words, $q1$ and $q2$ are the groups of Californians who entered and left quarantine on March 19th, 2020 and May 18th, 2020 respectively whereas $q3$ and $q4$ represent the population that followed the second stay at home order beginning on November 19, 2020 and ending on January 25, 2021.

The system of differential equations developed by Khan et al. (2020) is slightly modified and is as follows:

$$
\begin{aligned}
\frac{\partial S}{\partial t} &= -\beta S U^a - qS + \rho(E + R) \\
\frac{\partial U}{\partial t} &= -\beta S U^a - (q + \epsilon + \delta)U \\
\frac{\partial I}{\partial t} &= \delta U - (\gamma + \alpha)I \\
\frac{\partial R}{\partial t} &= \alpha I - \rho R \\
\frac{\partial D}{\partial t} &= \gamma I \\
\frac{\partial Q}{\partial t} &= q(U + S) \\
\frac{\partial E}{\partial t} &= \epsilon U - \rho E
\end{aligned}
$$

where

$$
\begin{aligned}
q = q1 \quad & \text{on Mar 19, 2020} \\
q = q2 \quad & \text{on May 18, 2020} \\
q = q3 \quad & \text{on Nov 19, 2020} \\
q = q4 \quad & \text{on Jan 25, 2021} \\
q = 0 \quad & \text{otherwise}
\end{aligned}
$$

To interpret the compartmental model, the proportion of individuals who are added to each state (S,Q,U,I,D,E,R) is governed by the top set of differential equations. The proportion of individuals in the Q (quarantine) state only changes on days when the four pulses $q1$, $q2$, $q3$ and $q4$ occur. The model is applied to daily reported case counts for SARS-COV-2 in California and minimizes the sum of squares between the actual and estimated detected cumulative detected infections (state I) as well as the predicted and true cumulative death toll from the virus (state D) as in Khan et al. (2020).

3.2.3 Model Evaluation

3.2.3.1 Fitting

In order to evaluate the fitting potential of both the HawkesN and SQUIDER comparison models, the measure of interest is the RMSE between predicted and actual doubling time during the training period. In this context, the reference date is defined as the first day taken into consideration when fitting both models while the cutoff date is the last day of the training period. When examining the fit for a given length training period > 5 days, we predict the cumulative and daily rate doubling times using just the first 5 days of the fitted model up to the day before the cutoff day. The resulting training RMSE is the square root of the mean of the squared differences between the estimated doubling times and the actual values.

For the HawkesN model, μ, κ and β are obtained using the modification of the least squares method originally described in Schoenberg (2021). Then, using the HawkesN Process Generator, 100 HawkesN branching processes are simulated where N is the population of California. The median of the 100 doubling times calculated for each fit is used as the estimate.

For the SQUIDER model, optimization is performed as described above. Then the estimated daily infections are calculated by dividing the reported infected state (I) by 14 to take into account the two weeks on average that an individual is usually in the infected state (Lewis et al., 2021). Here, the predicted doubling times used in determining the training RMSE are calculated using these case count estimates. Since some doubling times used to calculate the fitting RMSE extend into the forecasting period, the estimated doubling time is determined by continuing the disease process beyond the end of the training period while keeping fitted parameters β, δ, α, γ, ρ, ϵ, $U(0)$, a and q constant.

3.2.3.2 Forecasting

To examine the forecasting ability for both models, we use the RMSE between the actual and model estimated doubling times for both cumulative and daily rate. For context, both the HawkesN and SQUIDER models are fit from the reference date to the cutoff date. The cumulative doubling time is defined as the number of days post training period it takes to double the number of cumulative reported cases between the reference and cutoff days. Rate doubling time is the number of days after the training period it takes to double the number of daily cases detected on the cutoff date. For training periods of 10 to 31 days in length using the same starting reference date, each forecasted rate and cumulative doubling time is reported. The forecast RMSE, calculated individually for both doubling time metrics, is the average of the differences between the model predicted and true doubling times for all 22 fitting period lengths.

To forecast using the HawkesN model, 100 simulations are generated from the HawkesN Process Generator using fitted values of μ, κ and β as well as the population of California for N. The median cumulative and rate doubling times computed from the 100 simulations are used as the estimates when calculating the forecast RMSE for HawkesN.

For the SQUIDER model, after β, δ, α, γ, ρ, ϵ, $U(0)$, a and q are fit, they are held constant during the forecast period (Khan et al., 2020). The estimated daily and cumulative infections and deaths are also estimated at the end of the fitting period on the cutoff day. Then, the population proportions in each of the S,Q,U,I,D,E,R states are estimated each day thereafter with the fitted parameters constant. Predicted doubling times are calculated using the estimated cumulative and daily infection counts (state I) from the forecast period.

3.3 California SARS-COV-2 Data

The SARS-COV-2 coronavirus, originally discovered in December, 2019 in Wuhan, China (Centers for Disease Control and Prevention, 2021a), is the cause of the SARS-COV-2 disease outbreak originally declared a worldwide pandemic by the World Health Organization on March 11, 2020 (Cucinotta and Vanelli, 2020). Person to person transmission of the virus is usually airborne, causing symptoms ranging from high fever to body aches, loss of taste and smell and shortness of breath (Centers for Disease Control and Prevention, 2021a). While the vast majority of symptomatic SARS-COV-2 infections are considered mild, anyone is at risk of more severe symptoms requiring hospitalization, especially those who are immunocompromised or over age 65 (Centers for Disease Control and Prevention, 2021a). During peak surges, hospitals fill up and deaths increase as the number of people who are seriously or even critically ill rises rapidly (Centers for Disease Control and Prevention, 2021a). An accurate description of the spread of the virus during these surges and a widely used metric for this purpose is doubling time (Merler et al., 2013), (Kresin et al., 2021). Doubling time is easy to interpret and has been used to examine growth rates of various epidemics when control measures are either introduced or relaxed (Merler et al., 2013).

The California Department of Public Health (2021b) supplied official California statewide totals of the daily reported cases and deaths from SARS-COV-2 between February, 2020 and December, 2021. Daily reported cases are seen in Figure 3.1.

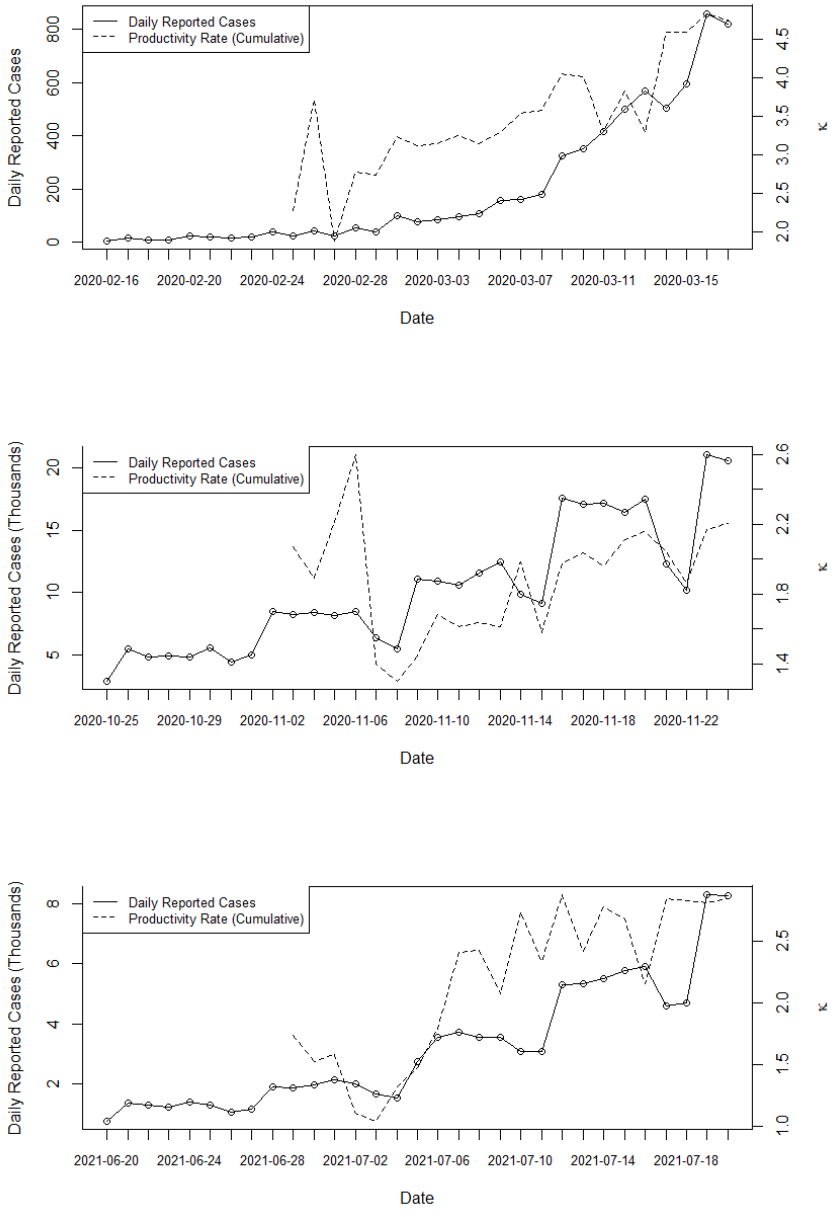


Figure 3.1: Daily number of reported cases and estimated productivity, $\hat{\kappa}$, for fitted HawkesN model, for Spring 2020 time period (top), Autumn 2020 (middle), Summer 2021 (bottom). For each day t, productivity is estimated using data from beginning of the plotted period up to and including day t.

We define each of three reference dates as the first day of the fitting period at the beginning of a given SARS-COV-2 surge. Each surge is analyzed separately, with cumulative cases and deaths set to zero within each surge.

The first time period of interest (beginning 2/16/20) is the initial Spring 2020 surge of SARS-COV-2 (California Department of Public Health, 2021b) when the population of California had not yet been exposed to the novel coronavirus as shown by a retrospective study of 1700 individuals with respiratory symptoms in December, 2019, none of whom had SARS-COV-2 (Hogan et al., 2020). Although testing for the virus was still quite limited in California during this time and there were severe testing problems (Patel, 2020), the reported Statewide count is used for the purposes of this analysis as the best option available.

The second surge analyzed (beginning 10/25/20) is during Autumn 2020, when there was another dramatic increase in hospitalizations and fatalities statewide (Centers for Disease Control and Prevention, 2021c). While there had been some population exposure due to the Spring 2020 surge and in Summer 2020, an insufficient proportion of the population was sufficiently immune to provide general herd immunity (Centers for Disease Control and Prevention, 2021c), (Fontanet and Cauchemez, 2020).

The third wave analyzed here (beginning 6/20/21) is during Summer 2021, when the more infectious Delta variant of SARS-COV-2 became dominant in California (Centers for Disease Control and Prevention, 2021b). By this time, vaccines effective against the Delta strain such as BNT162b2 developed by Pfizer (Bian et al., 2021) had been approved for emergency use by the U.S. Food and Drug Administration for those over the age of 12 (U.S. Food & Drug Administration, 2021). However, only 68% of the eligible population in California was fully

vaccinated by June 20, 2021 (California Department of Public Health, 2022) which did not provide for sufficient immunity against Delta.

3.4 Results

Figure 3.2 shows the median cumulative doubling time and daily rate doubling time for simulations of the HawkesN model governed by a range of κ and β values, each simulated 500 times, irrespective of the background rate μ.

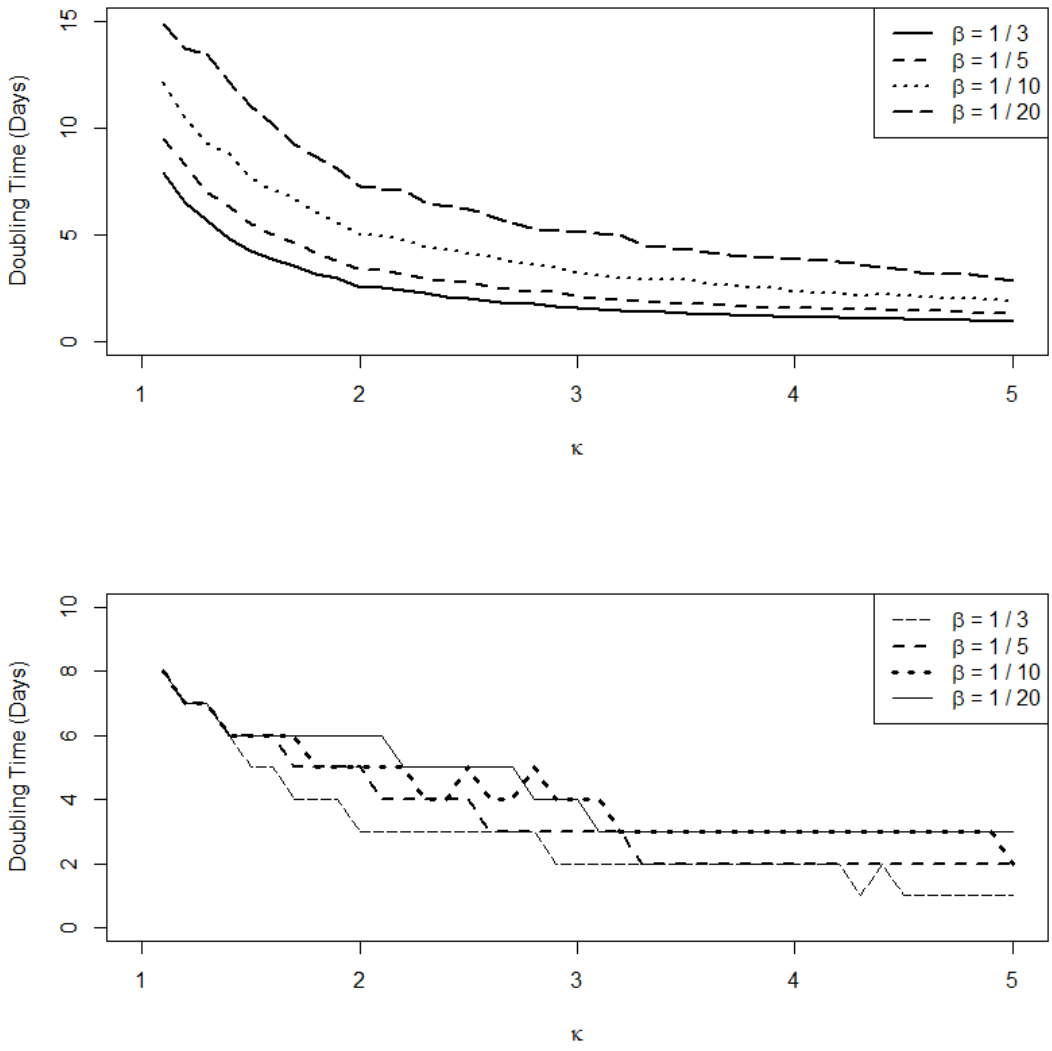


Figure 3.2: Doubling times over 500 simulations where $\beta = \frac{1}{3}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20}$. Top: median cumulative doubling time from the time when 50 total infections are observed. Bottom: Median daily rate doubling time from the time when 50 total infections are observed.

Here, β is the parametric estimate of the triggering function g where $\frac{1}{\beta}$ is the generational

length, or time between initial contagion and any resulting future infections. Both the cumulative doubling time and daily rate doubling time are defined as the time elapsed from the 50th recorded infection until the time of the corresponding doubling, as in the COVID-19 Dashboard provided by Johns Hopkins University and Medicine (2022). The median cumulative doubling times decrease as κ increases for each value of β. As β decreases, cumulative doubling times increase as expected, since β represents the inverse of generation length. However, as κ increases from 1.1 to 5.0, the effect of β on doubling time appears to dampen, for both the rate doubling times and cumulative doubling times.

Table 3.1 shows how the HawkesN parameter estimates, fit by least squares to the California SARS-COV-2 data, vary as the observation period varies.

Fitting Period (Days)	Spring 2020			Autumn 2020			Summer 2021		
—	μ	κ	β	μ	κ	β	μ	κ	β
10	10.4	2.3	0.09	3167.0	2.1	0.05	91.1	1.7	0.10
11	10.9	3.7	0.08	3468.7	1.9	0.05	95.8	1.5	0.05
12	10.2	1.9	0.10	3567.3	2.2	0.06	89.2	1.6	0.07
13	9.1	2.8	0.06	3183.9	2.6	0.03	90.6	1.1	0.07
14	9.3	2.7	0.05	3564.2	1.4	0.07	86.6	1.1	0.08
15	6.0	3.2	0.08	2466.5	1.3	0.10	88.2	1.3	0.04
16	4.4	3.1	0.08	2159.2	1.4	0.09	100.1	1.5	0.06
17	12.1	3.1	0.07	2408.2	1.7	0.06	100.7	1.8	0.06
18	8.3	3.2	0.07	2360.0	1.6	0.06	92.5	2.4	0.04
19	11.9	3.1	0.06	2434.2	1.6	0.06	109.0	2.4	0.04
20	10.1	3.3	0.06	2420.3	1.6	0.06	102.7	2.1	0.05
21	8.7	3.5	0.06	2222.0	2.0	0.05	105.4	2.7	0.05
22	10.0	3.6	0.06	2413.5	1.6	0.06	110.0	2.3	0.03
23	12.0	4.1	0.06	2402.1	2.0	0.05	119.1	2.9	0.04
24	10.0	4.0	0.06	2409.0	2.0	0.05	100.2	2.4	0.04
25	2.0	3.3	0.07	2402.4	2.0	0.05	101.2	2.8	0.04
26	8.0	3.8	0.06	2329.7	2.1	0.04	104.0	2.7	0.03
27	2.0	3.3	0.07	2182.7	2.2	0.04	102.3	2.2	0.04
28	18.0	4.6	0.05	2189.0	2.0	0.04	88.0	2.8	0.03
29	18.0	4.6	0.05	1862.3	1.9	0.05	104.1	2.8	0.04
30	17.9	4.8	0.05	2081.3	2.2	0.04	103.0	2.8	0.04
31	20.0	4.8	0.04	2366.0	2.2	0.04	102.7	2.8	0.04

Table 3.1: Fitted values for HawkesN parameters μ, κ, β, estimated for various length training periods starting on the reference date.

Within each of the three surges, the dataset's origin date is the same, and the length of the observation period is allowed to vary. The estimated background rate μ is highest for the Autumn 2020 time period, whereas the estimates of κ are generally highest during the initial Spring 2020 surge, when more of the population was susceptible and only minimal mitigation efforts were in place, in agreement with the justification for the recursive model in Schoenberg et al. (2019). Estimates of β are generally slightly higher during the initial

Spring 2020 surge but vary only minimally throughout, compared to the other HawkesN parameters due to a longer time period between exposure and symptomatic disease for the original variant (Centers for Disease Control and Prevention, 2021a).

Table 3.2 provides estimates for selected parameters (β, δ, γ, α) for the SQUIDER model fit to the same California SARS-COV-2 data.

Fit days	Spring 2020				Autumn 2020				Summer 2021			
—	β	δ	γ	α	β	δ	γ	α	β	δ	γ	α
10	0.696	0.473	6.639	0.337	0.999	0.128	7.077	0.092	0.885	0.828	6.142	0.298
11	0.703	0.526	5.587	0.275	0.792	0.457	6.889	0.001	0.841	0.502	6.149	0.303
12	0.689	0.491	4.870	0.339	0.681	0.471	6.647	0.110	0.696	0.522	6.107	0.244
13	0.705	0.501	4.165	0.321	0.919	0.830	6.447	0.073	0.825	0.759	6.046	0.097
14	0.691	0.492	3.623	0.315	0.707	0.570	6.266	0.048	0.529	0.275	5.974	0.215
15	0.682	0.480	3.134	0.114	0.865	0.840	6.155	0.136	0.726	0.607	5.890	0.210
16	0.699	0.490	2.642	0.283	0.998	0.903	6.567	0.727	0.819	0.104	5.802	0.001
17	0.732	0.509	2.696	0.299	0.804	0.531	5.914	0.961	0.937	0.318	5.698	0.129
18	0.705	0.490	2.922	0.297	0.891	0.579	5.799	0.054	0.915	0.793	5.575	0.132
19	0.678	0.474	2.908	0.261	0.689	0.608	5.705	0.091	0.911	0.667	5.443	0.247
20	0.712	0.489	3.083	0.282	0.726	0.504	5.629	0.043	0.731	0.683	5.311	0.161
21	0.695	0.484	3.040	0.272	0.672	0.631	5.565	0.024	0.952	0.695	5.194	0.233
22	0.719	0.496	2.903	0.297	0.687	0.679	5.517	0.070	0.270	0.134	5.089	0.019
23	0.717	0.495	3.081	0.284	0.836	0.792	5.496	0.001	0.729	0.675	4.974	0.141
24	0.721	0.486	3.076	0.344	0.999	0.947	5.345	0.141	0.727	0.698	4.839	0.092
25	0.755	0.509	3.017	0.302	0.999	0.899	5.688	0.094	0.778	0.746	4.700	0.045
26	0.704	0.467	2.805	0.294	0.802	0.671	5.735	0.192	0.856	0.318	4.567	0.002
27	0.686	0.467	2.656	0.269	0.984	0.990	5.256	0.045	0.685	0.506	4.509	0.291
28	0.720	0.470	2.460	0.457	0.770	0.432	5.592	0.001	0.533	0.480	4.325	0.091
29	0.721	0.480	2.379	0.353	0.802	0.421	5.339	0.001	0.999	0.955	4.120	0.511
30	0.771	0.528	2.464	0.363	0.590	0.484	5.324	0.001	0.744	0.663	4.101	0.218
31	0.694	0.468	2.637	0.272	0.944	0.585	5.168	0.001	0.522	0.385	3.996	0.044

Table 3.2: Fitted values for selected parameters of the SQUIDER model, α, β, δ and γ, estimated for various length training periods starting on the reference date.

The estimated contact rate β is relatively constant for most fit lengths in all three time periods, ranging between 0.6 to 0.9. The estimated testing rate δ is generally higher during the Autumn 2020 and Summer 2021 SARS-COV-2 surges than during Spring 2020. Also, the estimates of the fatality rate γ are generally highest in the Autumn 2020 time period, and correspondingly the estimated recovery rate α is lowest during Autumn 2020 as well.

In Figure 3.3, the simulated in-sample RMSEs for the HawkesN and SQUIDER models are compared, using cumulative doubling time as a metric.

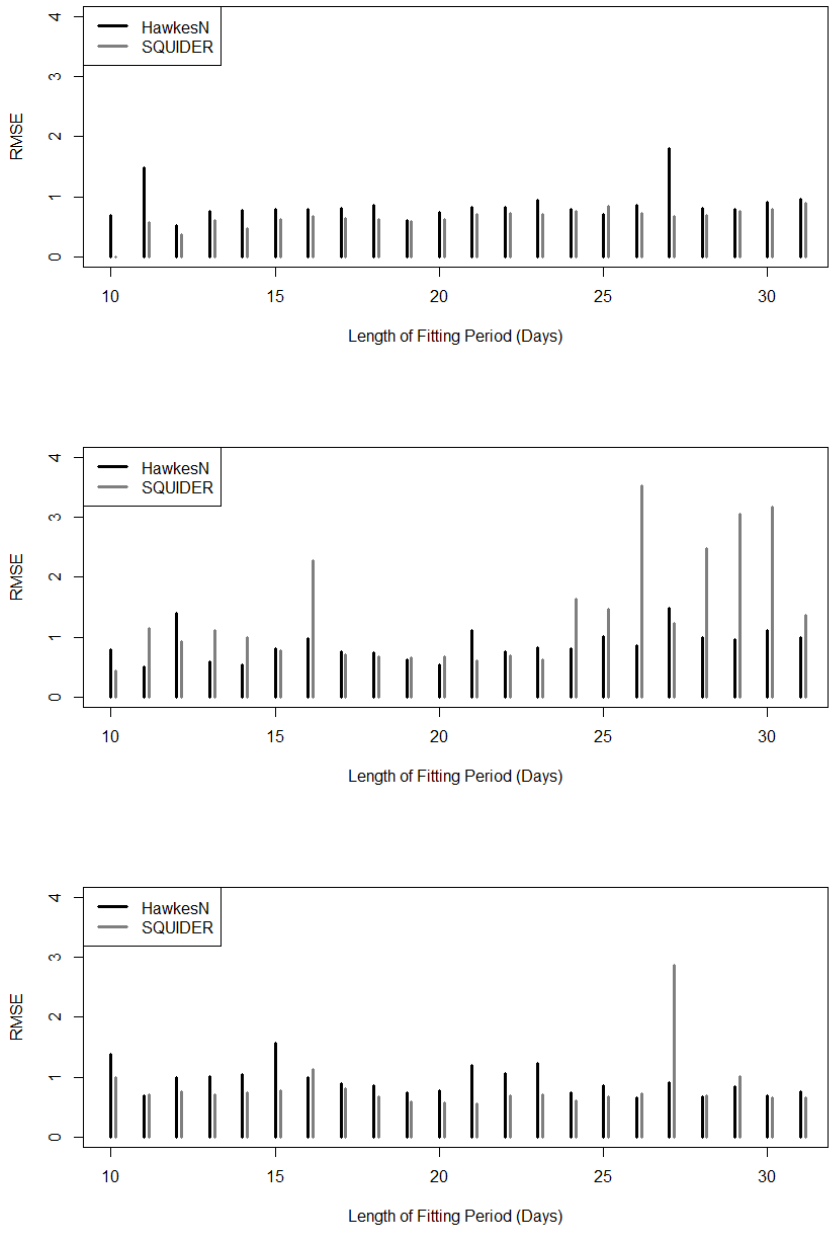


Figure 3.3: Fitting stage RMSEs of cumulative doubling time estimates for the HawkesN and SEIR models. Results for Spring 2020 time period (top), Autumn 2020 (middle), Summer 2021 (bottom).

The two models perform quite similarly throughout during the three in-sample periods, though for the Spring 2020 time period, the HawkesN model has a slightly higher training RMSE than does SQUIDER for most of the fitting period. For the Autumn 2020 data, the training RMSE is generally lower for HawkesN than for SQUIDER when the fitting period is 25 days or longer. For Summer 2021, the training RMSE is typically slightly higher for HawkesN than for SQUIDER. The higher in-sample RMSE for HawkesN for Spring 2020 and Summer 2021 is likely due to the fact that it is a simpler model with three parameters fit rather than twelve as in the SQUIDER model.

Figure 3.4 again compares the training RMSEs for the HawkesN and SQUIDER models during the same time periods, but uses daily rate doubling time as a metric.

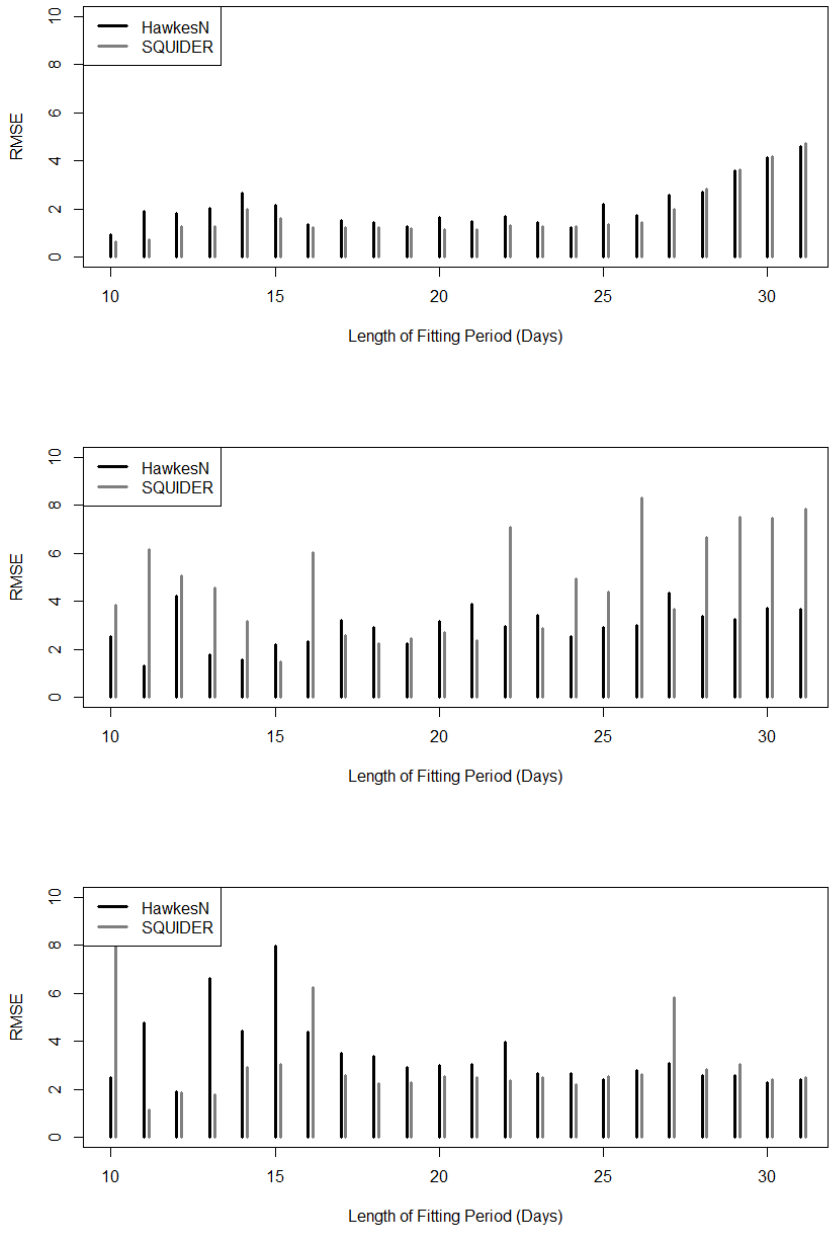


Figure 3.4: Fitting stage RMSEs of daily rate doubling time estimates for the HawkesN and SEIR models. Results for Spring 2020 time period (top), Autumn 2020 (middle), Summer 2021 (bottom).

Again the two models perform strikingly similarly. For Spring 2020 and Summer 2021, the fitting RMSE is slightly higher for the HawkesN model than for the SQUIDER model when the fitting duration is less than 26 days, while in Autumn 2020, the HawkesN model generally has a lower fitting RMSE than the SQUIDER model.

The results for each of the models forecasting doubling times, evaluated with different data from that used in the fitting of the models, are shown in Table 3.3 and in Figures 3.5 and 3.6 below. Table 3.3 reports the RMSE values when forecasting the cumulative doubling time and the daily rate doubling time for each surge.

Metric	Model	Spring 2020	Autumn 2020	Summer 2021
Cumulative	HawkesN	1.461	1.324	1.475
Cumulative	SQUIDER	1.087	2.637	1.552
Daily Cases	HawkesN	7.210	4.251	5.378
Daily Cases	SQUIDER	7.477	20.621	5.685

Table 3.3: RMSE for forecasted cumulative and daily rate doubling times for each model.

The out of sample RMSE is considerably lower for the HawkesN model than for the SQUIDER model overall, and especially for both the Autumn 2020 surge. During Autumn 2020, the HawkesN model has 49.8% smaller errors in forecasting cumulative doubling times, and 79.4% smaller errors in forecasting daily rate doubling times, compared to the SQUIDER model. For the initial Spring 2020 surge, however, the SQUIDER model has 26.9% smaller

errors in forecasting cumulative doubling times than the HawkesN model, although even in Spring 2020, the HawkesN model outperforms the SQUIDER model in forecasting daily rate doubling times.

Illustrated in Figure 3.5 is the forecasting performance of each model using the cumulative doubling time metric for each of the three time periods tested.

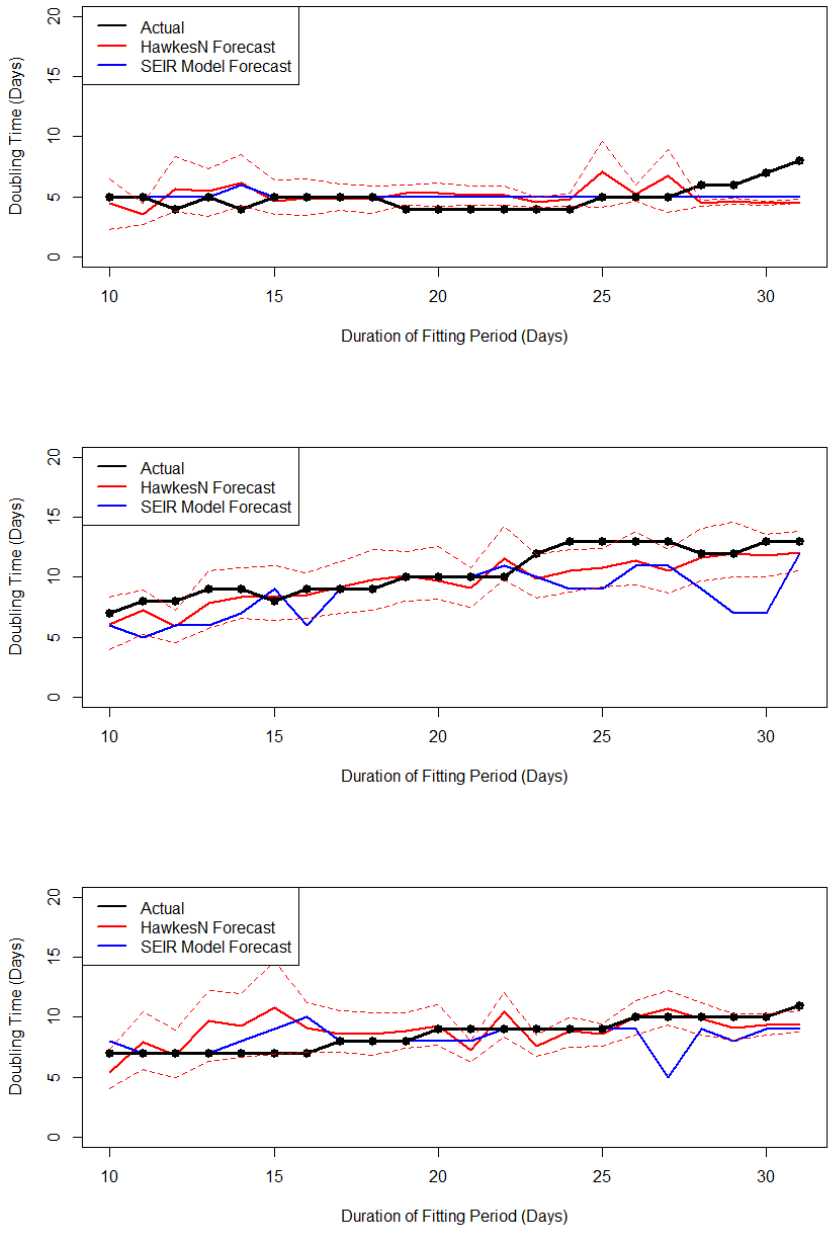


Figure 3.5: Cumulative doubling time estimates for HawkesN (red), SEIR (blue) and the actual cumulative doubling time (black) from the end of the fitting period. For HawkesN (red), 100 simulations run with the thick line representing the median and the thin dash lines representing the 90% bound based on the simulations. Results for Spring 2020 time period (top), Autumn 2020 (middle), Summer 2021 (bottom).

For the initial SARS-COV-2 surge in Spring 2020, both the SQUIDER model and the HawkesN model appear to slightly overestimate the cumulative doubling time when the fitting period is 19 to 24 days in length and considerably underestimate cumulative doubling times when the fitting period is 28 days or longer. The middle 90 percent ranges of HawkesN simulations contain the true cumulative doubling times in 10 out of 22 forecasts in Spring 2020, and are within one day of doing so in 8 other instances. For the Autumn 2020 COVID surge, while the SQUIDER model substantially underpredicts the cumulative doubling times particularly when the fitting period is 29-30 days, the HawkesN model appears to forecast accurately. The middle 90% of forecasted cumulative doubling times containing the observed cumulative doubling time in 18 of 22 forecasts. The HawkesN model forecasts also have higher accuracy than the SQUIDER model during the Summer 2021 SARS-COV-2 increase, with the SQUIDER model underestimating the cumulative doubling time when the fitting period is 25 days or longer. During Summer 2021, the middle 90% ranges of simulated cumulative doubling times for the HawkesN model contain the observed cumulative doubling times in 18 out of the 22 forecasts, and are within one day of doing so for all 22 forecasts in both Autumn 2020 and Summer 2021.

As shown in Figure 3.6, both the HawkesN and SQUIDER models forecast daily rate doubling times accurately in most cases.

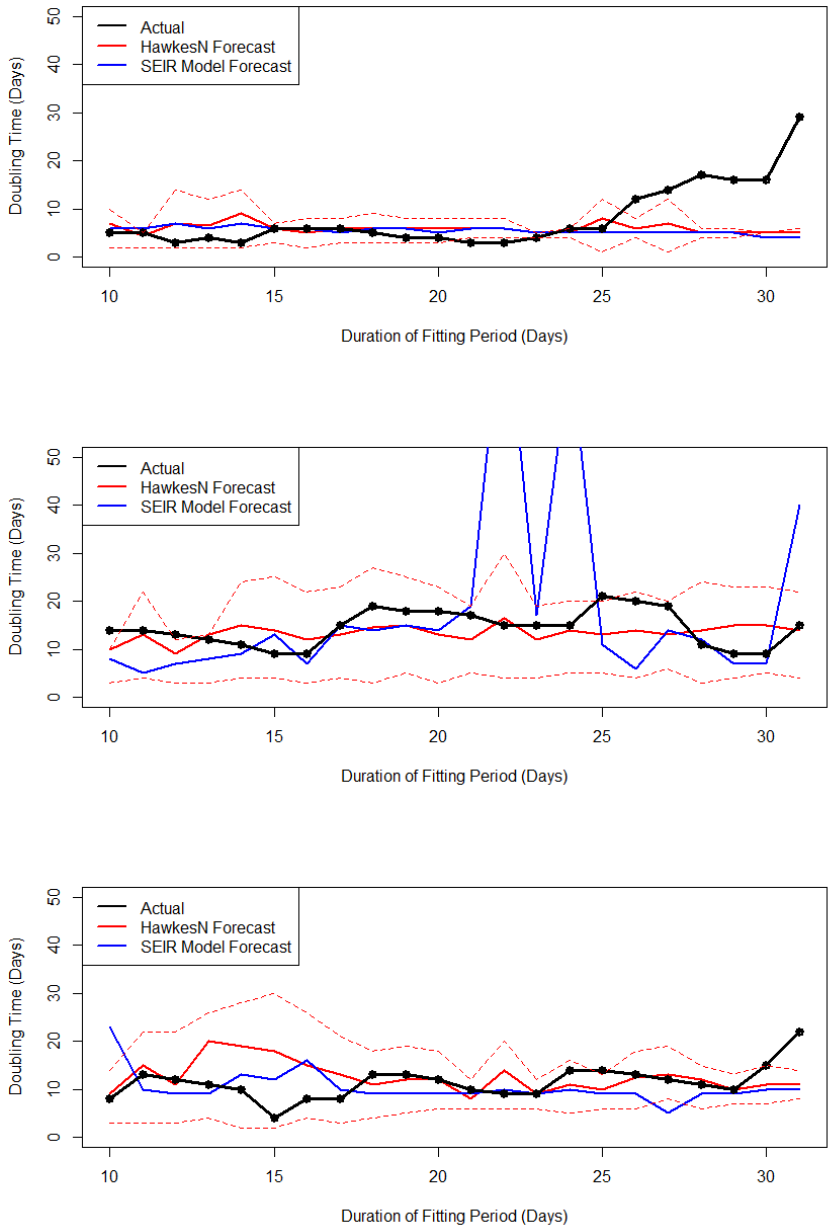


Figure 3.6: Daily rate doubling time estimates for HawkesN (red), SQUIDER (blue) and the actual daily rate doubling time (black) from the end of the fitting period. For HawkesN (red), 100 simulations run with the thick line representing the median and the thin dash lines representing the 90% bound based on the simulations. Results for Spring 2020 time period (top), Autumn 2020 (middle), Summer 2021 (bottom).

However, the SQUIDER forecasts of daily rate doubling times appear to be far more volatile and thus occasionally have much larger errors, particularly in Autumn 2020. Interestingly, both models underpredicted the rate doubling time in late March and early April 2020, when the rate of daily new recorded infections slowed, and neither model anticipated this trend. During both the Autumn 2020 and Spring 2021 timeframes, the HawkesN forecasts of rate doubling times had substantially smaller errors than the SQUIDER model. In total, the middle 90% range of simulations of HawkesN contains the true rate doubling time in 39 out of these 44 forecasts during these two time periods.

3.5 Conclusion

The HawkesN model appears to generally outperform the SQUIDER model in forecasting cumulative doubling times and rate doubling times during surges of SARS-COV-2 in California. The RMS errors in the HawkesN model when forecasting daily rate doubling times for all three SARS-COV-2 surges are lower than those of SQUIDER, as are those in forecasting cumulative doubling times during two of the three time periods of interest. The HawkesN model also appears to be substantially less prone to overfitting than the SQUIDER model as shown by the relative improvement compared to SQUIDER when tested on data not used in the fitting of the models.

An important item for future research involves exploring better ways to estimate the susceptible population when applying HawkesN to epidemic diseases such as SARS-COV-2 and its variants. The analysis here essentially assumes everyone in California is susceptible, and for N we use the State population according to the 2020 census (California Department of Public Health, 2021b). However, as noted in Rizoiu et al. (2018), estimating the size of the

susceptible population at any given point in time is not trivial and may perhaps be estimated by simulating a portion of the branching process first. Another potentially fruitful line of research may involve combining the HawkesN model and the recursive model, which could perhaps allow for a finite population as well as varying productivity, both shown to improve model performance individually when applied to forecasting epidemic diseases.

In addition, future research should explore whether the models assessed herein might be improved by taking into account vaccine uptake and waning immunity to the SARS-COV-2 virus as well as the impact of new variants (Centers for Disease Control and Prevention, 2021a). For this application, we have examined doubling times during three distinct periods, each of length 1-2 months, of sustained increase during which the population's immunity was unlikely to change substantially. However, a longer term forecast might need to take these extra factors into account. In the current formulation of the SQUIDER model, a subject is removed from the susceptible population if infected. Perhaps future formulations could remove individuals from the susceptible population when they take the recommended doses of vaccines such as BNT162b2 which was 95% effective against the original strain of SARS-COV-2 (Centers for Disease Control and Prevention, 2021a) and should be added back in when immunity wanes (Hamady et al., 2022) or a new variant reduces the effectiveness of such a vaccine (Centers for Disease Control and Prevention, 2021a). Estimating these quantities might be extremely difficult, however, and the results here suggest that, for estimating doubling times at least, the SQUIDER model may already be prone to overfitting, thus yielding larger errors compared to the simpler HawkesN model with fewer estimated parameters.

CHAPTER 4

Evaluation of HawkesN Model Using Least Squares Fitting Method in Estimating SARS-COV-2 in the United States

4.1 Introduction

As discussed in Chapter 3, the HawkesN model (Rizoiu et al., 2018) with an exponential kernel given by

$$\lambda(t) = (1 - \frac{N_t}{N}[\mu + \kappa \sum_{t_j<t} \phi(t - t_j)] \tag{4.1}$$

with kernel

$$\phi(u) = \theta e^{-\theta u} \tag{4.2}$$

is an adaptation of the Hawkes model which has been shown to take into account any non-stationarity when $\kappa > 1$ by including an additional term that assumes a finite population of susceptible individuals (Kresin et al., 2021). Traditionally, the parameters in the HawkesN model are fit using maximum likelihood estimation (MLE), allowing them to have useful properties such as asymptotic normality, efficiency and consistency (Ogata, 1978). However, the MLE method requires for the time of day of a given infection to be known, which is not usually recorded (Schoenberg, 2021). When only reported daily totals are known, a least squares estimation procedure proposed in Schoenberg (2021) has been shown to be effective. Applying the least squares method using HawkesN has been shown to successfully estimate

SARS-COV-2 cases in Indiana (Mohler et al., 2021) and nationally (Chiang et al., 2020). It has also been used to predict doubling time during three surges in California as in Chapter 3 and in Kaplan et al. (2022).

In this chapter, the effectiveness of HawkesN using least squares is further tested by expanding the size and scope of our analysis to fit and forecast daily case counts of SARS-COV-2 for all fifty U.S. states during three notable surges in Summer 2020, Winter 2020-2021 and Summer 2021. The predictive accuracy of HawkesN is compared to that of the SQUIDER (Susceptible, Quarantine, Undetected Infected, Infected, Dead, Exposed, Recovered) compartmental model (Khan et al., 2020) in a similar fashion as in Chapter 3 examining doubling times. A larger analysis as presented here allows us to evaluate the performance of the two models in forecasting many different disease spreading behaviors. As such, the purpose is to determine specific instances when HawkesN performs better than the compartmental model or vice-versa in order to guide further work.

4.2 Methods

4.2.1 Model Fitting

Here, we fit the HawkesN and SQUIDER models to five two-week periods ($5\times14=70$) of reported case counts in each state beginning on 5/1/2020, 10/1/2020, and 6/1/2021. For each model fit, we used the resulting parameter estimates to produce forecasts beyond the end of the fitting period.

4.2.1.1 The HawkesN Model

In fitting any Hawkes model (1.1), the values for the background rate μ, productivity κ, and the triggering function g must be estimated. This is usually done using maximum likelihood

estimation (Ogata, 1978), (Ogata, 1988), but when daily case counts are provided without time-of-day precision, the least squares technique introduced in Schoenberg (2021) works better as the time of day does not have to be randomly generated. While the least squares algorithm in Schoenberg (2021) and discussed in detail in Chapter 3 does well fitting case counts during the first 576 days of the COVID-19 pandemic in 50 states, the technique is slightly modified here to take into account the fact that the resulting parameters from this method are applied to out-of-sample forecasts. Notably, the fitting period is substantially shorter, including only five two-week periods for a total of 70 days. Also, the estimation of the productivity is of upmost importance in this case, since only the last value of κ during the fifth two week period is used in the ensuing forecast. The fitted values of the background rate μ and the triggering function g are constant for all five intervals.

While this chapter mainly focuses on the ability of HawkesN to forecast using the resulting parameter estimates from the least squares method, there is interest in evaluating the fit of the least squares algorithm in some situations. This is done using the same simulation method originally developed to fit HawkesN models in evaluating COVID-19 doubling time discussed in Chapter 3. The algorithm involves simulating a SEIR-Hawkes process, which takes advantage of the point process properties of the Hawkes family of models as well as the easy interpretability of the terms from the SIR compartmental models (Kresin et al., 2021).

4.2.2 Model Forecasting

4.2.2.1 HawkesN Model

To predict using the HawkesN model, we develop an algorithm to forecast HawkesN processes using fitted values of μ, κ and g obtained using least squares as well as the number of susceptibles from each state. In the case of SARS-COV-2, the susceptible group consists of the entire population due to the high infectiousness of the virus (Centers for Disease Control

and Prevention, 2021a). Rather than extend the branching process from the end of the fitting period, we predict the expected case counts using a simpler more direct method shown below.

First, we obtain the actual case counts a_{start} from the last interval of the fitting period with length T (two weeks in this application) and the fitted parameters from least squares. Second, we use the following iterative method to provide an estimate of the case count for a given day:

Step 1: Calculate the expected contribution β for the triggering component. In other words for $i \in 1, .., T$, the contribution is $\beta_i = g_i * a_{T-i+1}$. We denote a to be the estimated or actual case counts from the last T days.

Step 2: Estimate the conditional intensity λ for the given day by $\lambda_{T+1} = \mu + \kappa * \sum_{i=1}^{T} \beta_i$. Note κ is the productivity from the last interval of the fitting period.

Step 3: Estimate the case count for the current day by $a_{T+1} = Poisson(\lambda_{T+1})$.

Step 4: Update a to be the vector $a_2, ..., a_{T+1}$. Note that the old a_1 is removed from the a vector.

Running one iteration results in the estimate for the number of infections on a single day. To obtain a c-day forecast, simply iterate c times.

4.2.2.2 Model Evaluation

During the forecasting period, we use Root Mean Square Error (RMSE) to evaluate the accuracy of HawkesN and SQUIDER in predicting daily reported case counts. For HawkesN, the median estimated forecast from 100 simulations using the method above is applied when

calculating the RMSE. For SQUIDER, there is only one forecast simulation run since each of the fitted parameters is kept constant during the out-of-sample period.

With the purpose of seeing which model produces a better long-term forecast of daily case counts, we let i, $i \in [1, 40]$ be the forecast horizon (forecast length) and compare the prediction errors given by the two models iterating through i. In order to minimize prediction errors from any data irregularities (for example, missing values which are recorded as 0), we introduce the concept of lag, represented by a new starting point for the forecast j, $j \in [1, 30]$ days after the end of the fitting period. For each forecast horizon, i, we create a forecast for each value of j, thus resulting in a total of $30 \times 40 = 1200$ total forecasts for each state and surge.

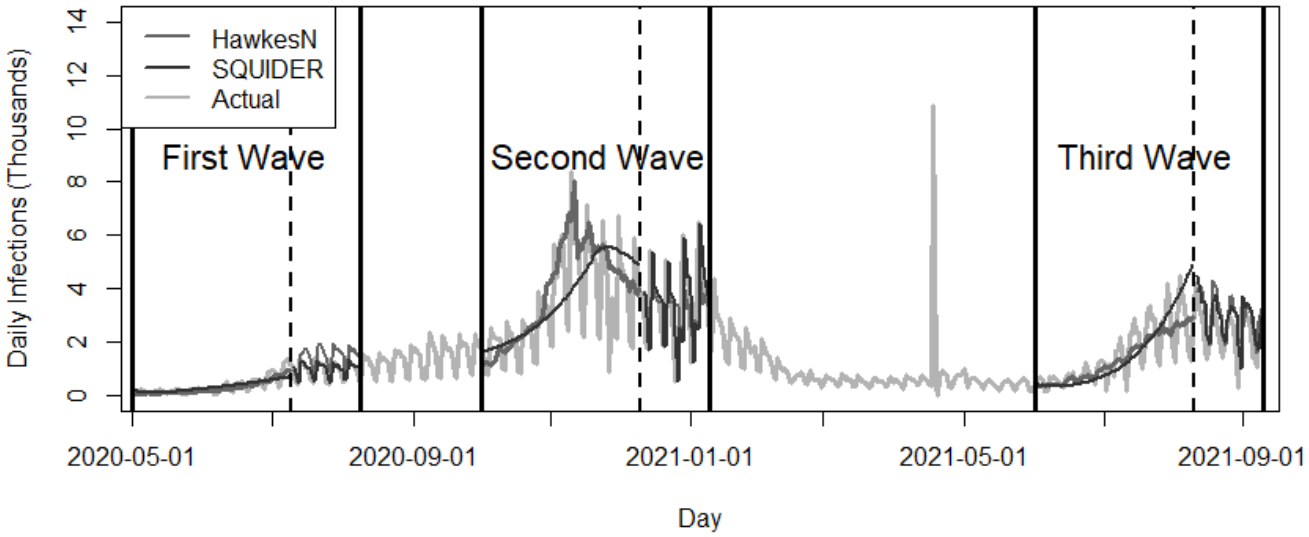
The total number of days for analysis of each surge is 70 fitting days + 70 forecasting days [1] = 140 total days.

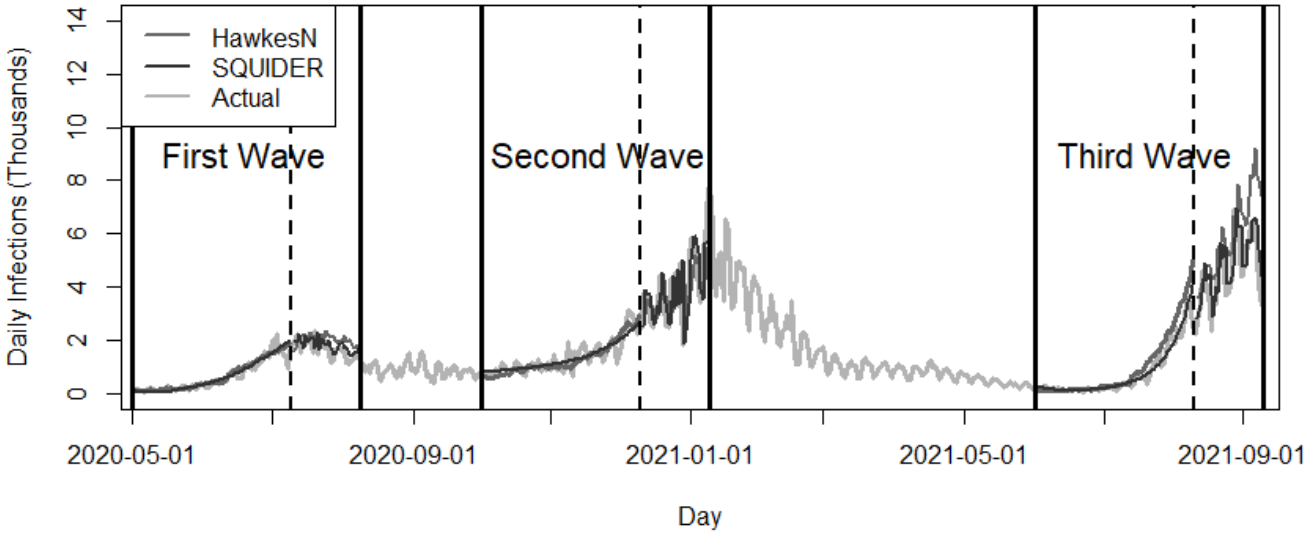
4.3 Nationwide SARS-COV-2 Data By State

As described in chapter 3, the SARS-COV-2 coronavirus, originally discovered in December, 2019 in Wuhan, China (Centers for Disease Control and Prevention, 2021a) is the cause of the COVID-19 pandemic declared by the World Health Organization on March 11, 2020 (Cucinotta and Vanelli, 2020). In Chapter 3, doubling time is used as the metric of interest in predicting virus spread during three surges in California. In this chapter, daily reported case counts of SARS-COV-2 are predicted in all fifty states during three notable periods when virus transmission has been at peak levels, provided by Centers for Disease Control and Prevention (2022).

[1] $max(i + j) = 70$

In Kaplan et al. (2022), doubling time is used as the metric of interest in predicting virus spread during three surges in California. Here, we evaluate forecasts of daily reported case counts of SARS-COV-2, using data provided by Centers for Disease Control and Prevention (2022), for each of the 50 states during three notable periods described below, when virus transmissions were at peak levels. This results in $3 \times 140 \times 50 = 21000$ observations of daily case counts in total, as there are three SARS-COV-2 surge periods considered, each consisting of 140 days, for each of the 50 states. These $21,000$ observed case counts are compared to forecasts for each of the two models considered, HawkesN and SQUIDER, and where each model is projected forward different numbers of days forward varying from 1 day to 40 days as described in the previous Section. Figure 4.1 shows the actual and fitted case counts as well as one-day forecasts for Missouri, South Carolina, Wisconsin and Oregon during all three surges.


HawkesN
SQUIDER
Actual
First Wave
Second Wave
Third Wave
Daily Infections (Thousands)
0
2
4
6
8
10
12
14
2020-05-01
2020-09-01
2021-01-01
2021-05-01
2021-09-01
Day


HawkesN
SQUIDER
Actual
First Wave
Second Wave
Third Wave
Daily Infections (Thousands)
0
2
4
6
8
10
12
14
2020-05-01
2020-09-01
2021-01-01
2021-05-01
2021-09-01
Day

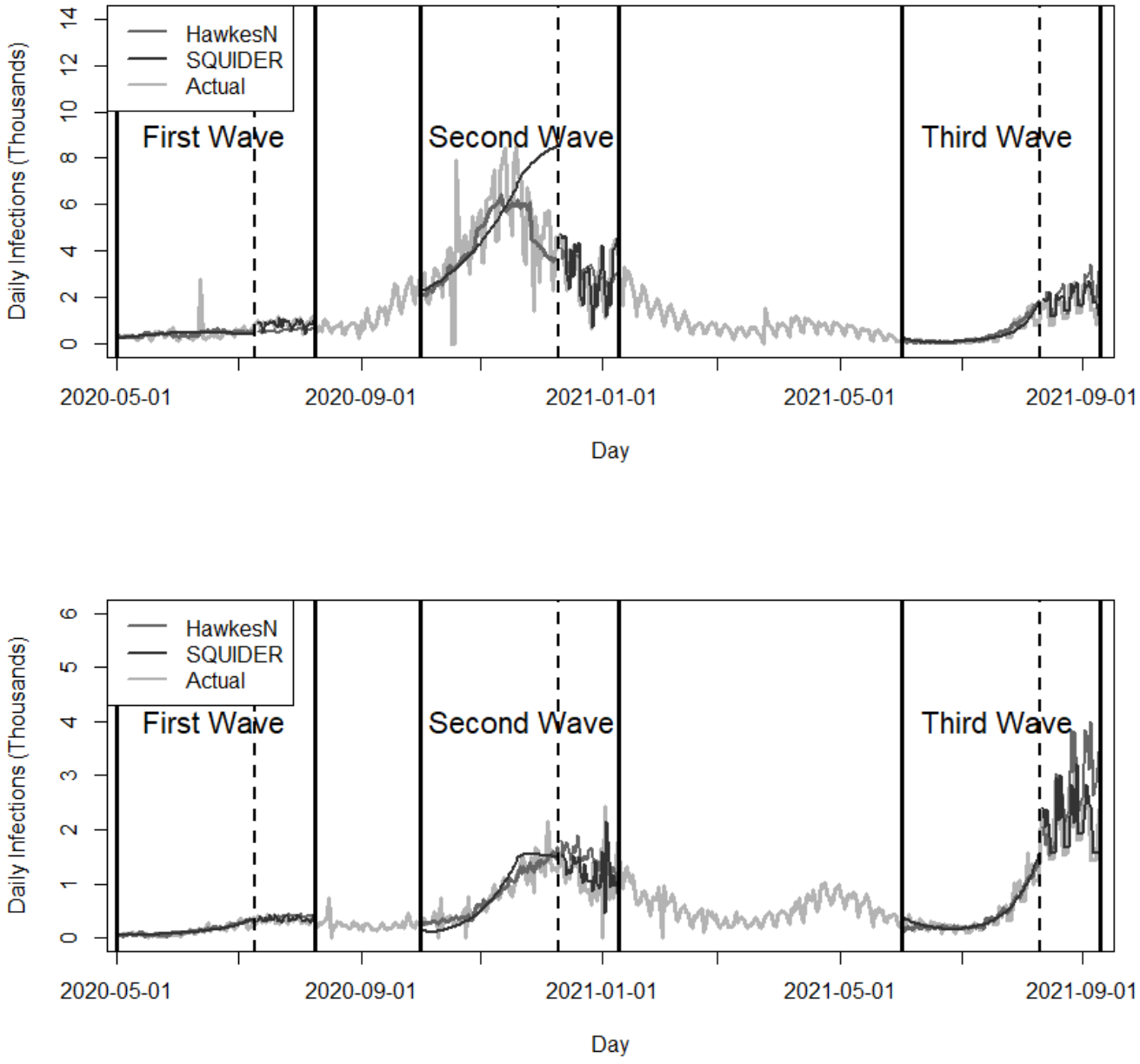


Figure 4.1: Fitted and actual reported case counts as well as one-day forecasts during three surges for Missouri (top), South Carolina (second from top), Wisconsin (second from bottom) and Oregon (bottom). Note that within each wave, to the left of the dotted line is the training period and to the right is the forecast period.

In this chapter, the first surge for all fifty states analyzed (beginning May 1, 2020) was in summer, 2020. The training period includes the first 70-days, that is, five consecutive two-week periods from May 1, 2020 to July 9, 2020 and the forecast period includes a second 70-days from July 10, 2020 to Sep 17, 2020. Although the virus initially began spreading in the United States in February, 2020, testing for the virus was still quite limited throughout the nation during the spring, so more reliable case counts began in earnest in summer, 2020 (Patel, 2020), (Centers for Disease Control and Prevention, 2021a). The summer 2020 surge in the United States was likely caused by the reopening of many businesses before there was sufficient immunity in the general population (Centers for Disease Control and Prevention, 2021a).

A second and larger surge of SARS-COV-2 (beginning Oct 1, 2020) occurred during autumn 2020 and winter 2021 throughout the United States (Centers for Disease Control and Prevention, 2021c). As mentioned in chapter 3, this resulted in a large increase in hospitalizations as well as the highest death toll during the pandemic to date (Centers for Disease Control and Prevention, 2021c). The fitting period includes the 70-days from Oct 1, 2020 to Dec 9, 2020 and the out-of-sample period is the next 70-days from Dec 10, 2020 to Feb 17, 2021. Even though vaccines effective against the initial strain of the virus had been approved for emergency use in December, 2020 for those over the age of 18 (U.S. Food & Drug Administration, 2021), virus transmission was already reaching a peak by that point and an insufficient proportion of the population was exposed to the virus to provide herd immunity (Centers for Disease Control and Prevention, 2021a), (Fontanet and Cauchemez, 2020).

A third notable surge (beginning June 1, 2021) occurred during summer 2021 when the

new and more transmissible Delta variant spread nationwide (Centers for Disease Control and Prevention, 2021b). The training period includes the five consecutive two-week periods from Jun 1, 2021 to Aug 9, 2021 and the forecast timeframe includes a second 70-days from Aug 10, 2021 to Oct 18, 2021. Despite the emergency use authorization of vaccines effective against SARS-COV-2 such as Pfizer's BNT162b2 and Moderna's mRNA-1273 for those over 18 years of age (U.S. Food & Drug Administration, 2021), insufficient vaccine induced immunity as well as the increased efficiency of the new Delta variant resulted in further spread of the SARS-COV-2 virus (Centers for Disease Control and Prevention, 2021b).

4.4 Results

4.4.1 Overall Model Comparison

Figure 4.2 shows the median absolute error between estimated and actual SARS-COV-2 case counts for the 150 HawkesN and 150 SQUIDER out-of-sample predictions given a forecast length of 1 through 40 days. Recall that for each HawkesN forecast, the median case count estimate from 100 simulations is used for calculating the error.

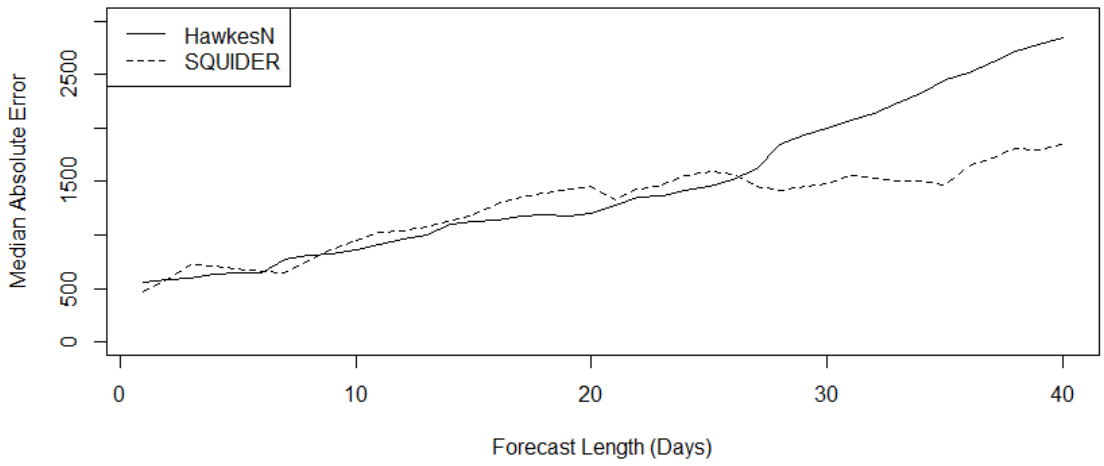


Figure 4.2: Median error between estimated and actual case counts for 150 HawkesN and 150 SQUIDER forecasts (50 states, 3 surges for each model). This is for forecasts of length 1 day up to 40 days inclusively.

As the length of the forecast increases, the absolute error for both HawkesN and SQUIDER rises due to a lack of certainty as expected with a longer range prediction. The median error is slightly lower for the HawkesN model as compared to the SQUIDER model for shorter term forecasts up to 4 weeks (28 days) in length, but is much higher for longer-term projections past 28 days.

In Figure 4.3, the RMS errors for each of the 150 HawkesN and SQUIDER forecasts of length 1 to 40 days are shown.

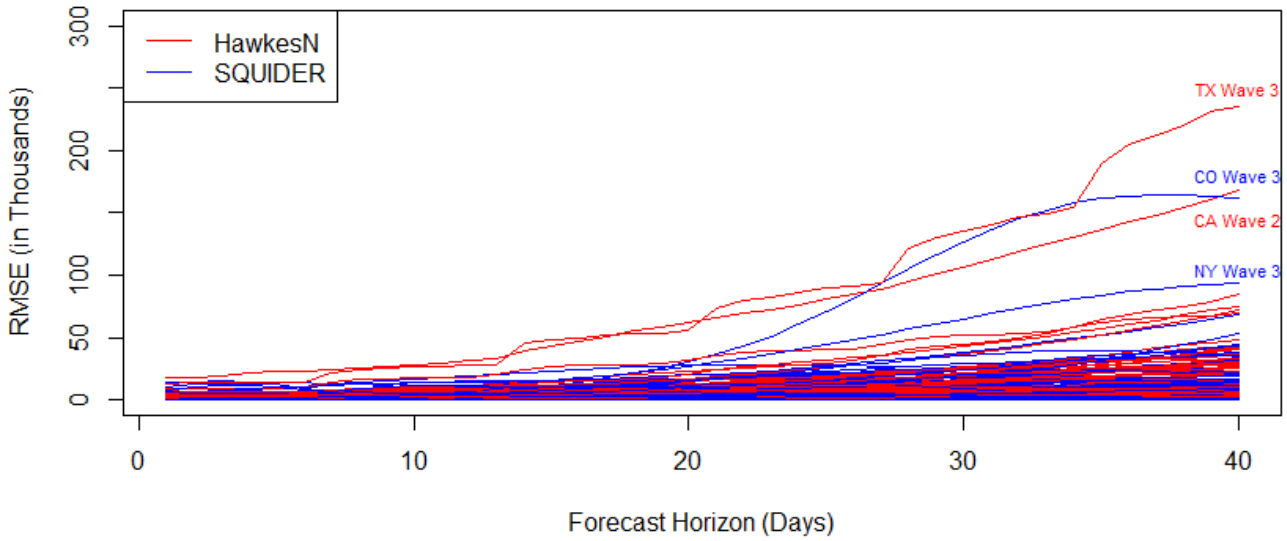


Figure 4.3: Root Mean Square Error (RMSE) between estimated and true case counts for each individual forecast for the HawkesN (red) and SQUIDER (blue) models for forecasts of length 1 day to 40 days.

As can be seen, the HawkesN forecasts for California during the autumn 2020 surge as well as Texas during the summer 2021 wave and the SQUIDER forecasts for Colorado and New York during the summer 2021 waves are flagged for having an abnormally high RMSE value relative to other forecasts. These are examined in greater detail in Figures 4.6 and 4.7.

4.4.2 Forecast Trends

As seen in Figure 4.4, we examine the effect of any observed dropoff from the apex of a given state's surge and the 40 day forecast performance between the two models. The peak week is calculated by finding the seven day period in the forecast period where the average

observed case count is highest. The final week count is the mean number of cases recorded during the last seven days of the forecast, or days 34 through 40.

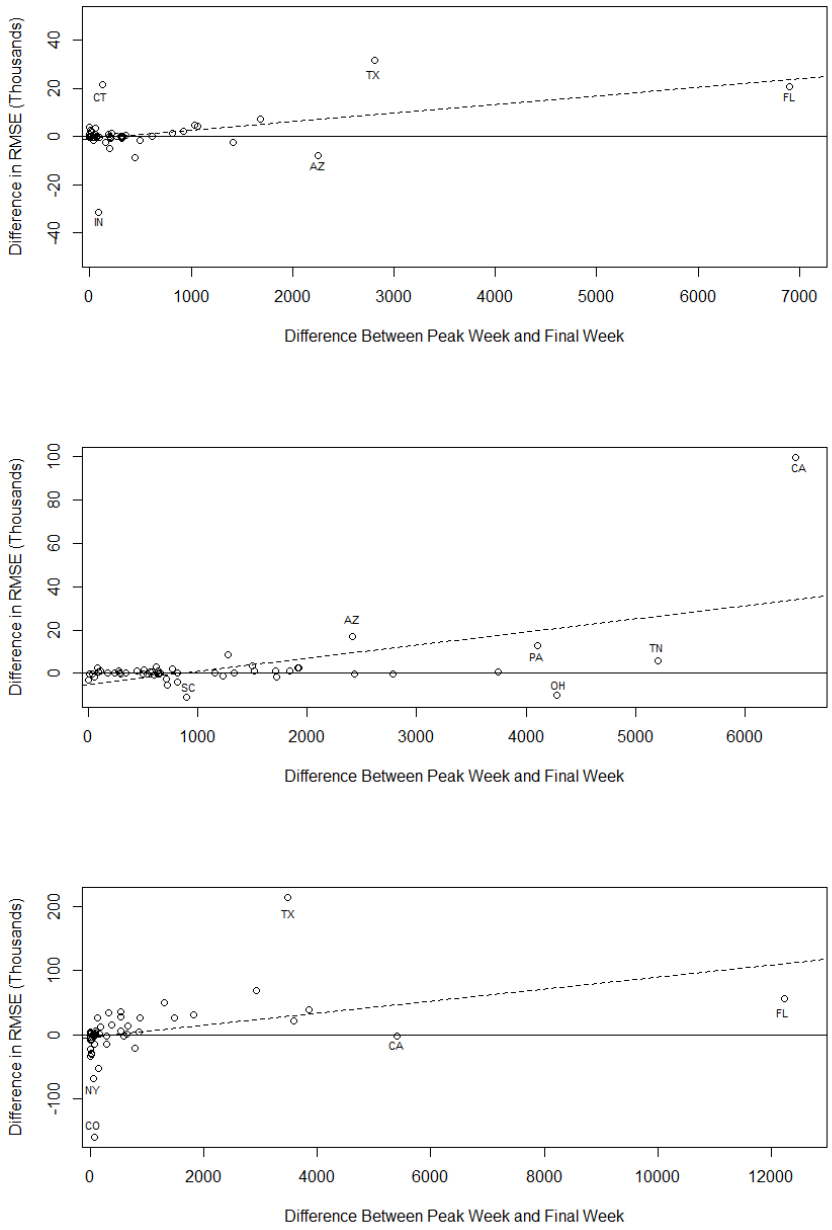


Figure 4.4: Scatterplot of observed decline in case count from the peak and the difference in Root Mean Square Error (RMSE) between the HawkesN and SQUIDER models for 40-day forecasts. This is for all 50 states during the summer 2020 surge (top), autumn 2020 surge (middle) and summer 2021 surge (bottom). The dotted line in each plot is the line of best fit.

For all three time periods of interest, the pattern is largely the same. The larger the drop, i.e. that the case count first increases rapidly followed by a peak and a sharp fall, the more the SQUIDER model outperforms the HawkesN model. Conversely, when the case count changes little, the HawkesN long term forecast performs better.

Another factor examined is whether forecasts are accurate for the HawkesN or SQUIDER model given an increase or decrease in the true reported productivity over time during the training period. Note that a change in the true reproduction rate implies a quadratic change in reported case counts. Figure 4.5 examines the relationship between the change in cases per day resulting from the quadratic coefficient over the course of the training period (x-axis) and the difference in the 40-day forecast RMSE between HawkesN and SQUIDER (y-axis).

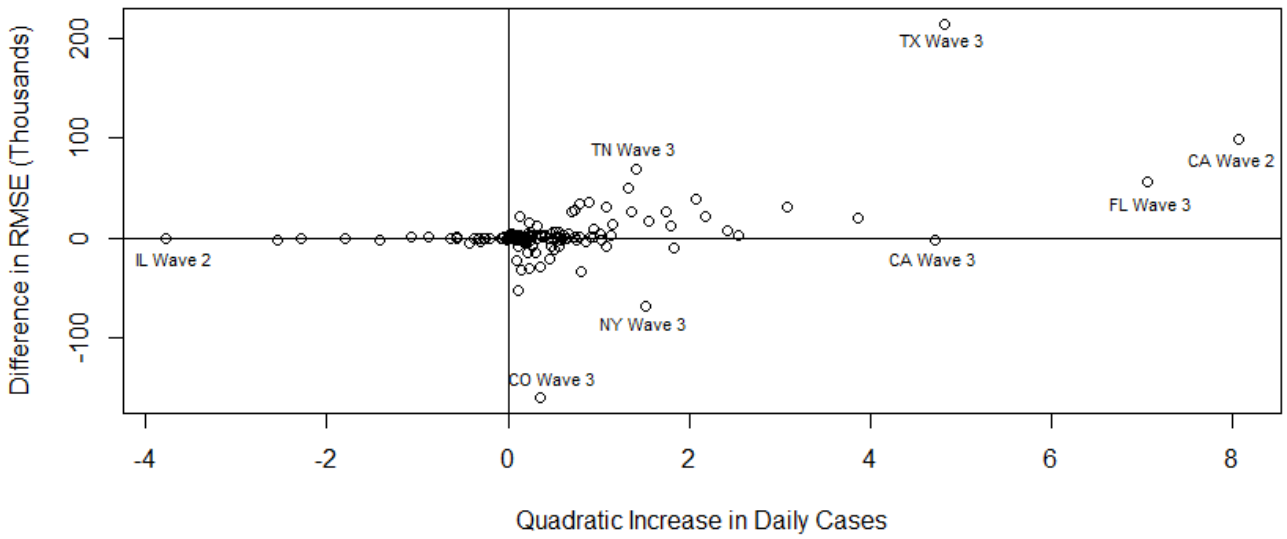


Figure 4.5: Scatterplot of the quadratic coefficient for cases per day during the training period (x-axis) and the difference in RMSE between the HawkesN and SQUIDER models for 40-day forecasts (y-axis). The x-axis represents the rate of change in productivity during the fitting period.

In the few cases where the productivity decreases over time, the two models largely perform similarly in forecasting. When productivity increases at a slower rate as seen in the summer 2021 surges in New York and Colorado (Figure 4.7), the HawkesN forecast is more likely to outperform the SQUIDER forecast than in scenarios where there is a faster rate of change such as in the California and Texas examples (Figure 4.6).

4.4.3 Specific Examples

In Figure 4.6, there are two cases where SQUIDER outperforms HawkesN. They are California during the autumn 2020 surge and Texas during the summer 2021 wave, also flagged in Figures 4.3 and 4.4.

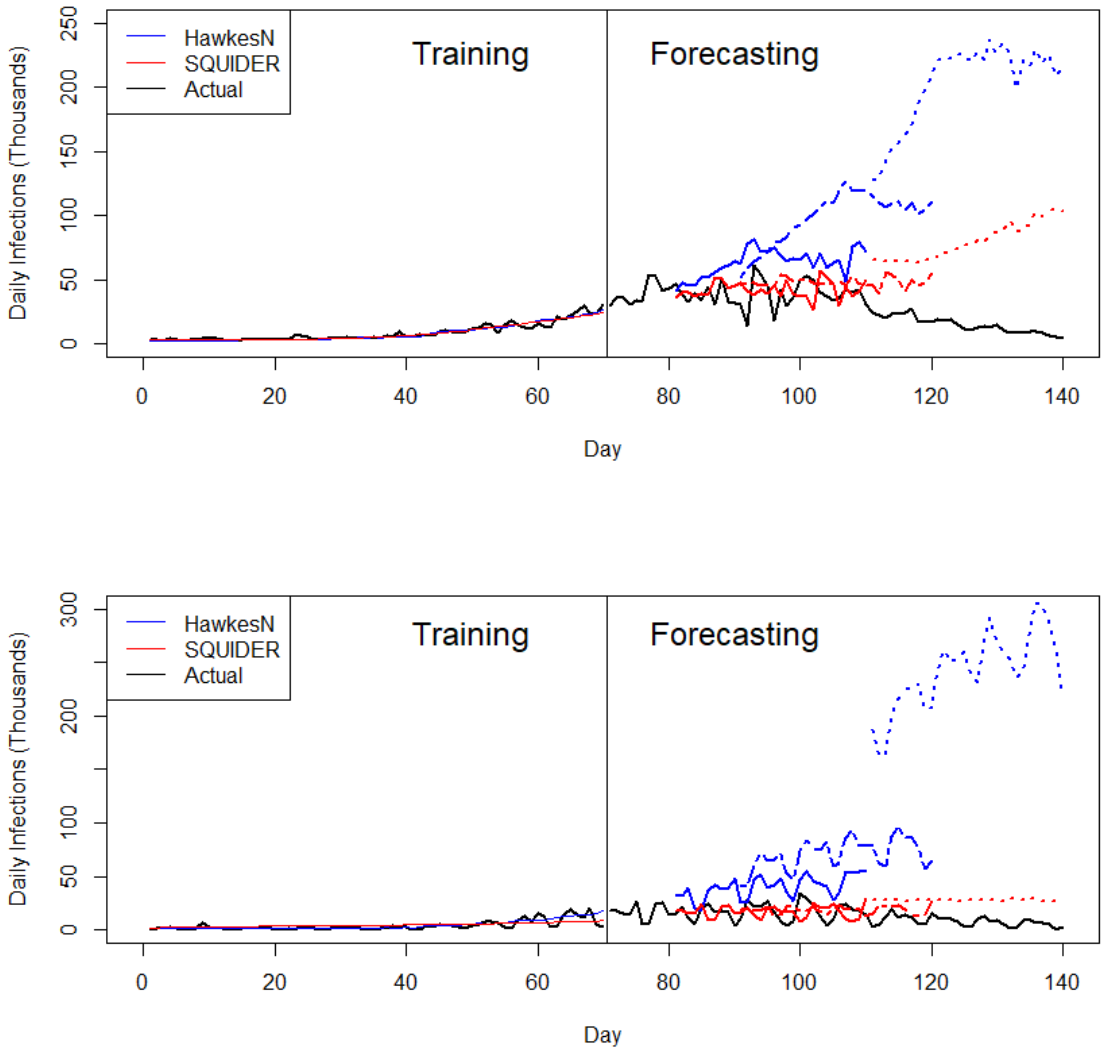


Figure 4.6: Fitting and forecasting for HawkesN (blue) and SQUIDER (red) for California during the autumn 2020 surge (top) and for Texas during the summer 2021 Delta variant wave (bottom). The dark blue and dark red lines represent 10 day forecasts, the medium hues represent 20 day forecasts and the lighter hues represent 40 day forecasts. Note that the Hawkes forecasts in blue contain the median and 95% bounds for 100 simulations run.

In the instance of the California surge, the estimated productivity during the last two weeks of the fitting period is similar between the HawkesN and SQUIDER models. However, while SQUIDER is able to adjust for the eventual peak in cases, the HawkesN model assumes a continued increase in the productivity rate during the forecast period leading to inaccurate predictions. For the summer 2021 Texas wave, the HawkesN model does account for the increasing productivity more successfully than the SQUIDER model during the last two weeks of the training period. However, HawkesN assumes that the rate of increase continues into the forecast period whereas SQUIDER does not, resulting in a more accurate forecast for the SQUIDER model.

Figure 4.7 shows two examples where the forecast provided by HawkesN is more accurate than that of the SQUIDER model, i.e. New York State during the summer 2021 wave and Colorado during the same surge.

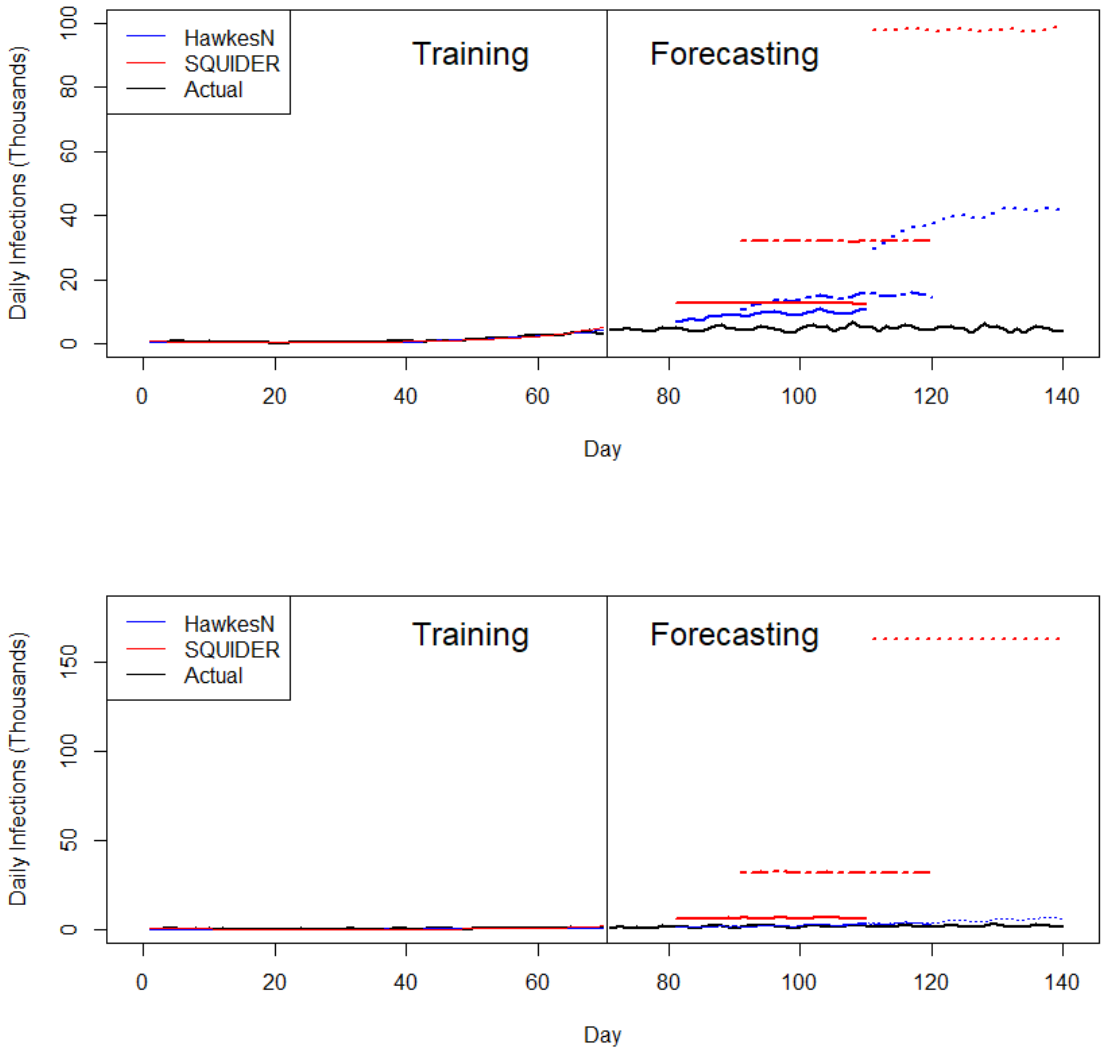


Figure 4.7: Fitting and forecasting for HawkesN (blue) and SQUIDER (red) for New York State during the summer 2021 surge (top) and for Colorado during the summer 2021 wave (bottom). The dark blue and dark red lines represent 10 day forecasts, the medium hues represent 20 day forecasts and the lighter hues represent 40 day forecasts. Note that the Hawkes forecasts in blue contain the median and 95% bounds for 100 simulations run.

In the New York State example, while there is a slight increase in cases during the last two weeks of the fitting period, it is not the rapid growth seen in the California and Texas scenarios shown in Figure 4.6. The HawkesN model predicts a modest bump in productivity during the forecast period whereas SEIR predicts a greater increase in cases when the true case count stays relatively stable in the out-of-sample period. Since the increase predicted by the SQUIDER model is larger, the forecast is less accurate than the one for HawkesN. In the Colorado scenario, the HawkesN model correctly predicts a relatively stable case count during the forecast period, whereas the SQUIDER model projects a sharp increase in cases.

4.5 Conclusion

We find mixed results when evaluating the forecasting errors of HawkesN and SQUIDER models. The HawkesN model typically outperforms SQUIDER in shorter term forecasting of less than 4 weeks, while SQUIDER is more accurate for longer term forecasts. In most scenarios when reported case counts increase rapidly followed by a sharp decline during the forecast period, the SQUIDER model outperforms the HawkesN model. On the other hand, HawkesN forecasts are more accurate when such changes are more modest. When case counts increase exponentially during the fitting period, HawkesN tends to overestimate the continued rate of spread more often than SQUIDER, which leads to overestimated case counts for some longer term forecasts.

An important item for further research involves improving the accuracy of the HawkesN in estimating the productivity at any given point in time. Here, we fit 5 two-week intervals with the value of κ obtained from the last interval held constant during the forecast period. However, as shown in the recursive model introduced in Schoenberg et al. (2019), it is possible to account for varying productivity over time. Future research should focus on ways

to combine the HawkesN (Rizoiu et al., 2018) and the recursive model (Schoenberg et al., 2019) to take into account a finite susceptible population as well as variable rate of change, both shown to individually improve model performance when used in epidemiological settings. Additionally, in Kaplan et al. (2021), a non-parametric version of the recursive model is introduced and successfully applied in forecasting cases of mumps, perhaps allowing for easy application when using the least squares algorithm where g is assumed to be a stepwise function.

In addition, it may be important to explore whether the HawkesN model could be improved by taking into account the impact of new variants and vaccine uptake (Centers for Disease Control and Prevention, 2021a). As shown in Rizoiu et al. (2018) and Centers for Disease Control and Prevention (2021a), when the susceptible population decreases due to vaccine induced or natural immunity, it becomes tougher for a given virus to spread at the same rate. Furthermore, as mentioned in chapter 3, the more infectious a new variant is, the shorter the doubling time . Also, a more infectious variant results in a steeper rise in cases followed by a short peak and dramatic fall when fewer susceptible individuals are left remaining (Rizoiu et al., 2018). An additional quadratic term controlling the change in productivity could perhaps be added to the HawkesN model, to predict when a peak should be expected based on the cumulative number of recorded cases as well as the estimated doubling time, so HawkesN might better account for steep declines in recorded cases immediately following surges.

CHAPTER 5

Closing Remarks

The motivation behind the research presented in this dissertation is to further improve upon the point process methodology used to fit and forecast communicable disease processes. While there are more widely used epidemiologic models such as the SIR (Susceptible, Infected, Recovered) compartmental set of equations (Kermack and McKendrick, 1927), the Hawkes point process model (Hawkes, 1971) has a self-exciting triggering component, which allows for any prior infection to trigger future cases, a key feature of such rapidly spreading diseases. An advantage for the point process model is that it does not require specifics of a given disease process to be known as is required for the compartmental model (see SQUIDER (Khan et al., 2020) and SVEILR (Li et al., 2018)). Rather, the triggering function g is estimated by taking into account time intervals between previously recorded events. As shown in Park et al. (2022), Meyer et al. (2012) and Meyer et al. (2017), even the basic Hawkes model outperforms compartmental models in forecasting communicable disease spread which is a clear indication that the newer point process application in epidemiology is worth pursuing further.

In chapter 2, a non-parametric version of the recursive adaptation of the Hawkes model (Schoenberg et al., 2019) is introduced and shown to outperform other point process and compartmental models in fitting and forecasting mumps in Pennsylvania. It is not only shown that a key principle of varying productivity is useful in enhancing predictive capability, but also that an expectation-maximization algorithm leading to a non-parametric fit can be used in such an application. However, the algorithm is computationally intensive as the matrix

required to generate triggering probabilities P_{ij} is an n^2 matrix. In the case of the mumps application, it takes 3 hours to run 100 iterations of the non-parametric algorithm. Another shortcoming is that small changes from iteration to iteration from such a large P_{ij} matrix often prevents complete convergence. However, the changes in practice are quite small and do not impact the overall fit of the model in a substantial way. Also, since the reported cases of mumps largely decreased over the years, there was no extra requirement for stationarity for $\kappa > 1$. Thus, the jury is still out as to whether the recursive model would be stable for an increasing epidemic. Future work could involve making the non-parametric algorithm more computationally friendly, given that the upper triangular part of the P_{ij} matrix where the probability of event i triggering event j given $t_i < t_j$ is 0. Another goal would be creating a stable version of the non-parametric recursive for an expanding epidemic.

The second part of the dissertation shows that there is promise in using point process models when case counts are increasing. It also discusses a newer method of fitting these models, allowing for a more precise estimate of fitted parameters. In chapter 3, it is shown that the HawkesN model (Rizoiu et al., 2018), which takes a finite susceptible population into account, outperforms the compartmental model in predicting doubling time in California. The HawkesN model in this scenario is fit using a least squares method, introduced in Schoenberg (2021). It is a much more practical method of fitting epidemiologic data, which often consists of daily reported infections rather than the time of day of an infection. In the analysis described in chapter 2, conducted before the least squares method was introduced, the exact time of infection is randomly chosen, whereas the exact time is not necessary when using least squares.

However, the results are mixed when the HawkesN model fitted using the least squares method is compared to the SQUIDER model forecasting case counts in the larger 50-state analysis. Further investigation is needed to determine what value of the productivity κ

should be used during the forecasting period. In chapter 4, the productivity value from the last two weeks of the training period is kept constant and is used in forecasting. Perhaps, a variable κ as proposed in the recursive model (Schoenberg et al., 2019) could be useful in improving such forecasting.

This dissertation presents cumulative work on several advancements in the use of point process models to fit and forecast communicable diseases. The overriding conclusion is that there are several properties of the self-exciting Hawkes model that can be used to more accurately describe and predict pathogen behavior over time. This exciting area in the field of statistics shows great promise and merits continued investigation and research.

Bibliography

Baddeley, A., Turner, R., Moller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:617–666.

Bahning, D., Rocchetti, I., Maruotti, A., and Hollinge, H. (2020). Estimating the undetected infections in the COVID-19 outbreak by harnessing capture-recapture methods. *Int J Infect Dis*, 97:197–201.

Balderama, E., Schoenberg, F., Murray, E., and Rundel, P. (2012). Application of branching point process models to the study of invasive red banana plants in Costa Rica. *JASA*, 107:467–476.

Bian, L., Gao, Q., Gao, F., Wang, Q., He, Q., Wu, X., Mao, Q., Xu, M., and Liang, Z. (2021). Impact of the Delta variant on vaccine efficacy and response strategies. *Expert Rev Vaccines*, pages 1–9.

Brémaud, P. (1981). Point Processes and Queues: Martingale Dynamics. *SpringerVerlag*.

California Department of Public Health (2021a). Limited stay at home order.

California Department of Public Health (2021b). COVID-19 Time-Series Metrics by County and State, California Open Data Portal. https://data.chhs.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state.

California Department of Public Health (2022). Vaccination data.

Centers for Disease Control and Prevention (2009). 2009 H1N1 Early Outbreak and Disease Characteristics.

Centers for Disease Control and Prevention (2011). Vaccine Coverage Levels - United States, 1962-2009.

Centers for Disease Control and Prevention (2019). *Epidemiology and Prevention of Vaccine-Preventable Diseases*.

Centers for Disease Control and Prevention (2021a). Coronavirus Disease 2019 (COVID-19).

Centers for Disease Control and Prevention (2021b). Delta Variant: What We Know About the Science.

Centers for Disease Control and Prevention (2021c). COVIDView - Key Updates for Week 3, ending January 30, 2021.

Centers for Disease Control and Prevention (2022). COVID data tracker.

Chiang, W., Liu, X., and Mohler, G. (2020). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *medRxiv*, doi.org/10.1101/2020.06.06.20124149.

Clements, R., Schoenberg, F., and Veen, A. (2012). Evaluation of space-time point process models using super-thinning. *Environmentrics*, 23:606–616.

Cucinotta, D. and Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Biomed*, 91:157–160.

Dubé, E., Vivion, M., and MacDonald, N. (2015). Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert Rev. Vaccines*, 14:99–117.

Elflein, J. (2019). New cases of mumps per 100,000 population in the U.S. from 1970 to 2017.

Executive Department State of California (2020). Executive Order N-33-20.

Fontanet, A. and Cauchemez, S. (2020). COVID-19 herd immunity: where are we? *Nat Rev Immunol.*, pages 1–2.

Gordon, J. (2017). nphawkes R package (Version 0.1.).

Gordon, J., Clements, R., Schoenberg, F., and Schorlemmer, D. (2015). Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. *Spatial Statistics*, 14b:133–150.

Hamady, A., Lee, J., and Loboda, Z. (2022). Waning antibody responses in COVID-19: what can we learn from the analysis of other coronaviruses? *Infection*, 50:11–25.

Hawkes, A. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90.

Hogan, C., Garamani, N., Sahoo, M., Huang, C., Zehnder, J., and Pinsky, B. (2020). Retrospective Screening for SARS-COV-2 RNA in California, USA, Late 2019. *Emerg Infect Dis*, 26:2487–2488.

Johns Hopkins University and Medicine (2022). Cumulative Cases by Days since 50th Confirmed Case.

Kaplan, A., Kresin, C., and Schoenberg, F. (sub 9/2022). Estimation of doubling time for SARS-COV-2 in California using HawkesN and SQUIDER models. *Journal of Infection*.

Kaplan, A., Park, J., Kresin, C., and Schoenberg, F. (2021). Nonparametric estimation of recursive point processes with application to mumps in Pennsylvania. *Biometrical Journal*, 64:20–32.

Kelly, D., Park, J., Harrigan, R., Hoff, N., Lee, S., Wannier, S., Selo, B., Massoko, M., Njoloko, B., Okitolonda-Wemakoy, E., Mbala-Kingebeni, P., Rutherford, G., Smith, T., Ahuka-Mundeke, S., Muyembe-Tamfum, J., Rimoin, A., and Schoenberg, F. (2019). Real-time predictions of the 2018-2019 Ebola virus disease outbreak in the Democratic Republic of Congo using Hawkes point process models. *Epidemics*, 28.

Kermack, W. and McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115:700–721.

Khan, Z., Van Bussel, F., and Hussain, F. (2020). A predictive model for COVID-19 spread – with application to eight US states and how to end the pandemic. *Epidemiology & Infection*, 148.

Kirchner, M. (2016). Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 28:2494–2525.

Kirchner, M. (2017). An estimation procedure for the Hawkes process. *Quant Financ*, 17:571–595.

Kirchner, M. and Bercher, A. (2018). A nonparametric estimation procedure for the Hawkes process comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation*, 88:1106–1116.

Kong, Q., M.A., R., and Xie, L. (2020). Modeling information cascades with self-exciting processes via generalized epidemic models.

Kresin, C., Schoenberg, F., and Mohler, G. (2021). Comparison of Hawkes and SEIR models for the spread of COVID-19. *Advances and Applications in Statistics*, 74:83–106.

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1:1–20.

Lewis, M., Duca, L., Marcenac, P., Dietrich, E., Gregory, C., Fields, V., Banks, M., Rispens, J., Hall, A., Harcourt, J., Tamin, A., Willardson, S., Kiphibane, T., Christensen, K., Dunn, A., Tate, J., Nabity, S., Matanock, A., and Kirking, H. (2021). Characteristics and Timing of Initial Virus Shedding in Severe Acute Respiratory Syndrome Coronavirus 2, Utah, USA. *Emerging Infectious Diseases*, 27.

Lewis, P. and Shelder, G. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26:403–413.

Li, Y., Liu, X., and Wang, L. (2018). Modelling the transmission dynamics and control of mumps in mainland China. *Int J Environ Res Public Health*, 15.

Marsan, D. and Lengliné, O. (2008). Extending Earthquakes' Reach Through Cascading. *Science*, 319.

Merler, S., Ajelli, M., Fumanelli, L., and Vespignani, A. (2013). Containing the accidental laboratory escape of potential pandemic influenza viruses. *BMC Medicine*, 11.

Meyer, S., Elias, J., and Hohle, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*.

Meyer, S., L., H., and Hohle, M. (2017). twinstim: An endemic-epidemic modeling framework for spatio-temporal point patterns. *Journal of Statistical Software*, 77.

Mohler, G. (2014). Marked point process hotspots maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, 30.

Mohler, G., Schoenberg, F., Short, M., and Sledge, D. (2021). Analyzing the impacts of public policy on COVID-19 transmission: a case study of the role of model and dataset selection using data from Indiana. *Statistics and Public Policy*, 8:1–8.

Mohler, G., Short, M., Brantingham, P., Schoenberg, F., and Tita, G. (2011). Self-exciting point process modeling of crime. *JASA*, 106:100–108.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30:243–261.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83:9–27.

Ogata, Y. (1998). Space-time point process models for earthquake occurrences. *Ann. Inst. Statist. Math*, 50:379–402.

Park, J., Chaffee, A., Harrigan, R., and F.P., S. (2022). A non-parametric Hawkes model of the spread of Ebola in West Africa. *J. Appl. Stat*, 49:621–637.

Park, J., Schoenberg, F., Brantingham, P., and Bertozzi, A. (2021). Investigating clustering and violence interruption in gang-related violent crime data using spatial-temporal point processes with covariates. *JASA*, 116:1674–1687.

Patel, N. (2020). Why the CDC botched its coronavirus testing. *MIT Technology Review*.

Rizoiu, M., Mishra, S., Kong, Q., Carman, M., and Xie, L. (2018). SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Processes. *Proceedings of the 2018 World Wide Web Conference*, pages 419–428.

Schoenberg, F. (2013). Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America*, 103:601–605.

Schoenberg, F. (2017). Introduction to Point Processes.

Schoenberg, F. (sub 9/2021). Estimating COVID-19 transmission time using Hawkes point processes. *Annals of Applied Statistics*.

Schoenberg, F., Hoffmann, M., and Harrigan, R. (2019). A recursive point process model for infectious diseases. *AISM*, 71:1271–1287.

Shmueli, G. and Lichtendahl, K. (2016). *Practical Time Series Forecasting with R: A Hands-On Guide*. Axelrod Schnall, 2 edition.

Stoyan, D. and Grabarnik, P. (1991). Second order characteristics for stochastic strutures connected with Gibbs point processes. *Mathematische Nachrichten*, 151:95–100.

U.S. Food & Drug Administration (May 10, 2021). Coronavirus (covid-19) update: FDA Authorizes Pfizer-BioNTech COVID-19 Vaccine for Emergency Use in Adolescents in Another Important Action in Fight Against Pandemic.

Van Panhuis, W., Cross, A., and Burke, D. (2018). Counts of Mumps reported in UNITED STATES OF AMERICA: 1923-2017 (version 2.0, april 1, 2018): Project Tycho data release.

Veen, A. and Schoenberg, F. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *JASA*, 103:614–624.

Wheatley, S., Filimonov, V., and Sornette, D. (2014). Estimation of the Hawkes process with renewal immigration using the EM algorithm. *Swiss Finance Institute Research Paper*, pages 14–53.

Yang, Y., Etesami, J., and Kiyavash, N. (2017). Online learning for multivariate Hawkes processes. *Advances in Neural Information Processing Systems*, page 4938–4947.

Zhuang, J. and Mateu, J. (2019). A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society Series A*, 182:919–942.