Title
A comparison between HawkesN and SIR-family models in forecasting COVID-19

Permalink
https://escholarship.org/uc/item/9gc9k7x7

Author
Farzin-Nia, Sasha

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A comparison between HawkesN and SIR-family models in forecasting COVID-19

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Sasha Farzin-Nia

2022

ABSTRACT OF THE THESIS

A comparison between HawkesN and SIR-family models in forecasting COVID-19

by

Sasha Farzin-Nia

Master of Applied Statistics

University of California, Los Angeles, 2022

Professor Frederick R. Paik Schoenberg, Chair

Starting late in 2019 in the Wuhan province of China, COVID-19 has become a pandemic in the last few years. We use mathematical and statistical modeling in the hope of proving that self-exciting models perform better than SIR-family models when it comes to modeling in Epidemiology. We take a look at two models, the SQUIDER compartmental differential equation model, and also the HawkesN self exciting process model. For training the SQUIDER model, we minimize the Sum of Square Errors between actual and predicted cumulative cases and deaths, and use weights to account for the cumulative cases, and cumulative deaths, respectively. In forecasting, we use fitted parameter values and the true case counts at the end of each lag, and forecast forward 3-days, and take the RMSE between the estimate forecast and the actual case counts of COVID-19. We fit the HawkesN model by using a Least Squares algorithm to fit the data we have and use an algorithm proposed in another paper to simulate the HawkesN method. Similar to the forecasting of the SQUIDER model, we take 10 3-day forecasting periods with a lag of 1 day through 10 days, and then calculate the RMSE. Taking a look at data from the CDC website, we fit and forecast for Oregon, South Carolina, and Washington State, and indeed see that except for a very few cases, overall the HawkesN process performs a lot better than the

SQUIDER model in forecasting the daily cases of COVID-19.

The thesis of Sasha Farzin-Nia is approved.

Mark S. Handcock

Nicolas Christou

Frederick R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2022

I would like to dedicate this thesis to my wonderful parents, whom without, I would have not been able to pave this path that I am currently following. Thank you for always motivating me, when things seem to become too much to handle. I love you both.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

x

CHAPTER 1

Introduction

First detected in the Wuhan province of China, late in 2019, COVID-19 is a disease which is caused by the SARS-CoV-2 virus [4]. It has since spread throughout the planet and become a pandemic. One of the questions we would like to try and answer as Statisticians is *'Can we predict the number of daily number of cases of COVID-19?'* Of course, this is a million dollar question, for if we could, with certainty answer this question, lock-downs, vaccination requirements, and other forms of minimizing the number of cases could more effectively be placed, and the pandemic could be put to an end. Whilst it is not possible to get a definitive answer, we can use mathematical and statistical modeling methods to emulate a fairly accurate representation of the timeline of daily cases. This paper will take a look at two model frameworks, compartmental models and point-process models. With the Root Mean Squared Error (RMSE) as our metric, our aim will be to compare the two in their ability to forecast the daily cases of COVID-19. One of the most common types of compartmental models is known as the SIR (Susceptible $\rightarrow$ Infected $\rightarrow$ Recovered) differential equation model [20]. Such model frameworks have been used frequently in the past for epidemiology, due to its ability to simplify the task at hand and allow modelers to predict key features of a disease merely by estimating certain parameters [20]. Examples of infectious diseases which SIR models and their extensions have been used are SARS [12] and Ebola [15]. This leads us to the next model framework, point-processes. This paper will focus on Hawkes processes and its extension, HawkesN, which have been shown to be a better fit when it comes to fitting and forecasting [8]. Such models have been used to forecast the number of cases of Chlamydia [18], Ebola [15], and the original coronavirus, SARS [21]. Aside from modeling infectious diseases, Hawkes models and its extensions have been proven to be the best models to

be used in the case of earthquakes [13].

CHAPTER 2

Data

Our data was collected from the CDC COVID Data Tracker website [5], where we downloaded datasets for North Carolina, Oregon, South Carolina, and Washington. For each of the States, we downloaded four datasets, containing the data on the number of Daily Cases, the number of Daily Deaths, the number of Cumulative Cases, and the number of Cumulative Deaths. Using the *dplyr* package in R, we merged the datasets so we could have one true data source for our modeling algorithms. For the SIR-family model, we also needed the population for each of the states. Using the following web-scraping packages in R, *readxl*, *rvest*, and *readr*, we read in the data from wikipedia [22].

CHAPTER 3

Theoretical Background

3.1 Compartmental Models

SIR models and its variants are what are known as compartmental models and are used frequently to model and forecast the number of infections caused by a certain infectious disease, such as the COVID-19 Pandemic [9]. This models types' name is attributed to the fact that out of a total population you are working with, N, we can split this population into different compartments. From [20], at its simplest form we can split the population into 3 different components as follows:

3.1.1 SIR Models

- Susceptibles, $S(t)$ - These are the members of the population, which are susceptible to being infected by the disease of interest;

- Infected, $I(t)$ - These are the members of the population, which have been infected by the disease;

- Recovered/Removed, $R(t)$ - These are the members of the population, which have been removed from the model system, due to one of two reasons:

  1. They have **recovered**, and thus left the infected state, or;

  2. They have died, and thus are **removed**, from the model.

The 3 compartments, are all functions of time, t.

In this simplest case, we make the assumption that we start with a certain number of the population which are susceptible, which will then at some point, move to the infected state, and from which will be moved to the recovered/removed state [20].

The SIR models and its variants are considered to be closed system models [20], that is:

$$S(t) + I(t) + R(t) = N, \qquad (3.1)$$

where N, is a constant.
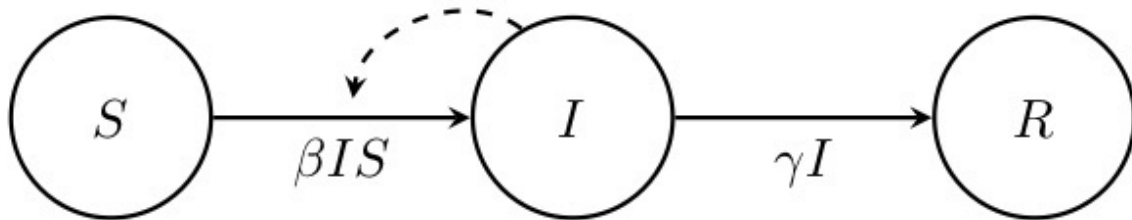
Figure 3.1: Flow diagram of Susceptible $\rightarrow$ Infected $\rightarrow$ Recovered Compartmental Model [3].

We want to take a look at flow of transmission between each individual component. In order to do this we take a look at the different possible options that can happen at each compartment [3]:

At S: From the susceptible state, we can only go to the infected state [1]. Since we are starting

---

[1]This is only the case for the simplest SIR model.

with a number of people in the state moving to another state, we have a decrease, thus a negative sign will be required, and we have

$$\frac{dS}{dt} = -\alpha SI,$$

where α is known as the transmission coefficient, and is calculated by $\alpha = pq$, where p is the proportion of the infected population in which the susceptible is in contact with, and q is the proportion of contacts with the infected population a susceptible has, that actually causes infection.

We will now skip to the R compartment, as it is simple to envision - from the infected state, we can accept an influx of the population, and thus we are adding to the population, requiring a positive sign, as follows [3]:

$$\frac{dR}{dt} = \beta I$$

where β is known as the recovery coefficient, and is calculated as follows $\beta = \frac{1}{k}$, where k is the average number of days which the infection lasts in the individual.

At I: We can accept an influx of people who have come from the susceptible state, requiring a positive sign, and we can also move from the infected state to the recovered/removed state, requiring a negative sign.

Mathematically, this can be confirmed. Since from equation 3.1, we have that:

$$S(t) + I(t) + R(t) = N,$$

and thus:

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$$
$$\Rightarrow -\alpha SI + \frac{dI}{dt} + \beta I = 0$$
$$\Rightarrow \frac{dI}{dt} = \alpha SI - \beta I,$$

6

which confirms our reasoning above.

In summary, we end up with the following system of equations:

$$\begin{cases} \dfrac{dS}{dt} = -\alpha SI \\[2mm] \dfrac{dI}{dt} = \alpha SI - \beta I \\[2mm] \dfrac{dR}{dt} = \beta I \end{cases} \tag{3.2}$$

3.2 SQUIDER Model

With the COVID-19 pandemic, comes increased complexities, which in order to model and fore-cast more accurately, means that we have to increase the complexity of the model, and build around the simplistic SIR model. Our model is based off of the Cambridge article, 'A predictive model for Covid-19 spread – with application to eight US states and how to end the pandemic' [9].

In order to code our algorithm we start by using initial parameters based off of the article, and use optimization to fine tune them. The initial parameters are set out as follows in table **??**.

While with the SIR model, we have 3 compartments, in this thesis, we will be taking a look at a model which contains 7 different components as mentioned in [9], which are shown as follows:

- Susceptibles, S(t) - these are the members of the population who are susceptible to getting infected by the disease;

- Social Distanced, Q(t) - these are the members of the population who adhere to the methods set by the World Health Organization (WHO) [14] in order to try and decrease the number of infections, protect themselves, and those around them;

- Undetected Infected, U(t) - Since this virus, has the ability to attach itself to its host, but not necessarily cause symptoms to present (these members of the population are known to

Parameter	Value
a	1
β	0.7
δ	0.5
ϵ	0.1
γ	0.02
ρ	0.01
α	0.3
U	1×10^{-5}
q1	0.1
q2	-0.5
q3	0.1
q4	-0.5

Table 3.1: Initial parameter values for the SQUIDER Model - [9].

be **asymptomatic** [4]), but can infect others around them;

- Detected Infected, I(t) - these are the members of the population who are infected with COVID-19 and are showing symptoms;

- Detected Deaths, D(t) - these are the members of the population who we have recorded to have died;

- Undetected Recovery/Death, E(t) - these are the members of the population who have either recovered or died without taking account of;

- Detected Recovered, R(t) - these are the members of the population who have been monitored and have made a full recovery. In this model, we assume they are able to return back to the susceptible state [9].

Similar to the SIR, the assumption is that this is a closed system model [9], that is:

$$S(t) + Q(t) + U(t) + I(t) + D(t) + E(t) + R(t) = N$$

Without loss of generality, from [9], the rate equations are thus as follows:

$$\begin{cases} \dfrac{dS}{dt} = \rho(E + R) - \beta SU^\alpha - q(t)S \\[2mm] \dfrac{dQ}{dt} = q(t)(U + S) \\[2mm] \dfrac{dU}{dt} = \beta SU^\alpha - (q(t) + \epsilon + \delta)U \\[2mm] \dfrac{dI}{dt} = \delta U - (\gamma + \alpha)I \\[2mm] \dfrac{dD}{dt} = \gamma I \\[2mm] \dfrac{dE}{dt} = \epsilon U - \rho E \\[2mm] \dfrac{dR}{dt} = \alpha I - \rho R \end{cases} \tag{3.3}$$

3.3   Point Processes

Point processes take place in nature, and the data is thus used for a number of scientific applications, including seismology [13], epidemiology [15], and finance [2]. A point process, ϕ, is a random collection of points that live in a measure space, S. We represent this mathematically as follows [6]:

$$\phi = \{X_i\}, \ i \in \mathbb{N}$$

If considering purely temporal (relating to time) data, S, the space in which the different event points fall, are a portion of the real number line [17]. For ease of readability, we will focus solely on temporal point processes from here on forth. We can describe such processes as a counting

process, $N(t)$ [17]. $N(t)$ is defined as the number of points occurring at or before time, t in the interval, $[0, T]$. It is important to note that the resulting process has to be non-decreasing, right-continuous, and may only take non-negative integer values [17].

3.3.1 Modeling with Point Processes

In modeling a point process, we define it with its conditional probability, $\lambda(t)$ [17]. This represents the infinitesimal expected rate at which points are accumulating at time t, given information on all points occurring prior to time t. Mathematically, we define the conditional intensity, as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}[N(t + \Delta t) - N(t) | \mathcal{H}(t)]}{\Delta t}, \qquad (3.4)$$

where $\mathcal{H}$ is the history of the arrivals up to time, t [17].

A point process, $N(t)$, in which the covariance, $\text{cov}\{N(s, t), N(t, u)\} > 0$ for $s < t < u$, is known to be self-exciting [17]. We can think of a self-exciting point process as a stochastic domino effect, where if one point occurs in the self-exciting process, other points are more likely to occur also. For the purpose of this thesis, we consider a 'point' to be a single infection from the SARS-CoV-2 virus.

Self-exiting point processes are used to model events in which the points are temporally clustered, meaning that the arrival of a point increases the likelihood of observing similar points in the near future [17]. Some typical use-cases for self-exciting point processes are when modeling seismic activity, or financial activity such as the panic selling of stocks [1].

3.4 Hawkes Processes

Named after mathematician Alan G. Hawkes, the Hawkes point process model is a common self-exciting point process [7].

Consider a scenario, using $\{\tau_1, \tau_2, \cdot, \tau_n\}$ to represent an observed sequence of past arrival times of the Hawkes process up to the time, t. It has a conditional intensity defined as follows [17]:

$$\lambda(t) = \nu(t) + \kappa \sum_{i:\tau_i < t} g(t - \tau_i), \tag{3.5}$$

where $\nu(t)$ represents the predetermined background rate, and κ is the infection reproduction rate.

3.4.1 HawkesN Processes

As described in [16], the addition of a population of size N, further controls the branching process compared to the standard Hawkes process. For the purpose of this thesis, the main difference between a normal Hawkes process and a HawkesN process is that once an individual is infected, they are excluded from reinfection, which is shown mathematically in [16], as follows:

$$\lambda(t) = \left(1 - \frac{N_t}{N}\right)\left(\nu(t) + \kappa \sum_{i:\tau_i < t} g(t - \tau_i)\right), \tag{3.6}$$

where N_t represents the number of individuals which have been infected, and thus are no longer susceptible, at time t.

HawkesN has been shown to model epidemiologic branching processes effectively, as described in [12]. Moving forward, we will be discussing the results obtained by using the SQUIDER Model and the HawkesN Model. The methodology will be discussed as such in the next section.

11

CHAPTER 4

Methodology

4.1 SIR - Extension Model

4.1.1 Optimization of Model Parameters

We fit our model to a series of datasets constructed from web scraping the following websites:

- Wikipedia - List of U.S. States and Territories by Population [22]

- Center for Disease Control and Prevention - COVID Data Tracker [5].

Our goal was to forecast the number of reported daily cases of COVID-19. In order to do this, we needed to optimize the set of differential equations in the model by minimizing the Sum of Square Errors (SSE) between actual and predicted cumulative cases and deaths. We used a weighted approach to calculating the SSE, where one weight took into account the cumulative cases, and one weight took into account the cumulative deaths.

Using [9], as mentioned previously, we found that the system of differential equations

governing the SQUIDER model is:

$$\begin{cases} \dfrac{dS}{dt} = -\beta SU^\alpha - qS + \rho\left(E + R\right) \\[2mm] \dfrac{dU}{dt} = -\beta SU^\alpha - \left(q + \epsilon + \delta\right)U \\[2mm] \dfrac{dI}{dt} = \delta U - \left(\gamma + \alpha\right)I \\[2mm] \dfrac{dR}{dt} = \alpha I - \rho R \\[2mm] \dfrac{dD}{dt} = \gamma I \\[2mm] \dfrac{dQ}{dt} = q\left(U + S\right) \\[2mm] \dfrac{dE}{dt} = \epsilon U - \rho E \end{cases} \tag{4.1}$$

where the following are the rates of:

- α: recovery

- β: transmission

- δ: testing

- ϵ: undetected outcomes

- γ: known deaths

- ρ: waning immunity.

4.1.2 Forecasting Methods

For this forecasting section, we first begin by introducing the topic of lag. This can be most aptly defined as the number of days after the end of the fitting period in which an n-day forecasting period begins at. For forecasting the SQUIDER model we used the fitted values for each parameter $(\alpha, \beta, \delta, \epsilon, \gamma, \rho, q1, q2, q3, q4)$ as well as the true case count at the end of the lag and forecasted forward 3-days, resulting in a forecasting estimate for daily case counts of COVID-19. With this, we calculated the RMSE between the forecast estimate and the actual case count of COVID-19.

13

4.2 HawkesN Model

4.2.1 Least Squares Algorithm

The HawkesN model was fitted to a series of training datasets from the data constructed by web scraping the CDC website as mentioned in Section 2, with the intention of predicting the number of reported daily cases of COVID-19. In order to model using HawkesN, the first thing to do was to estimate the parameters.

Given the data we have, since we had daily counts rather than specific times of infection, instead of using the usual method of Maximum Likelihood Estimation, we used a proposed Non-Parametric approach to fitting the data [19], using Least Squares Estimation which is based on the relationship between HawkesN processes and autoregressive processes detailed in [10] [11]. The HawkesN parameters were fit to the reported daily case counts, for the first N days, for COVID-19 in Oregon, South Carolina, and Washington. Mathematically, this is represented as follows [19]:

$$\sum_{t=1}^{N} \left(N(t) - \left[\mu + \sum_{i=1}^{14} \kappa(t-i)g(i)N(t-i) \right] \right) \tag{4.2}$$

Since we fitted five 14-day periods, this resulted in us having five kappa parameters and 14 g parameters, however using one constant background rate, μ. This resulted in an optimization of parameters based around κ and g, thus not optimizing μ as well as it could. To overcome this, we minimized the sum of squares shown above using a range of values of μ, and selected the μ which resulted in the lowest RMSE value.

4.2.2 HawkesN Simulator

The algorithm for simulating the HawkesN method is proposed in [12].

4.2.3 Forecasting Methods

For forecasting using HawkesN, we took ten 3-day forecasting periods with a lag of 1 day through 10 days, and calculated the RMSE value between the actual daily case counts and the forecasted daily case counts of COVID-19 for each individual lag.

CHAPTER 5

Results
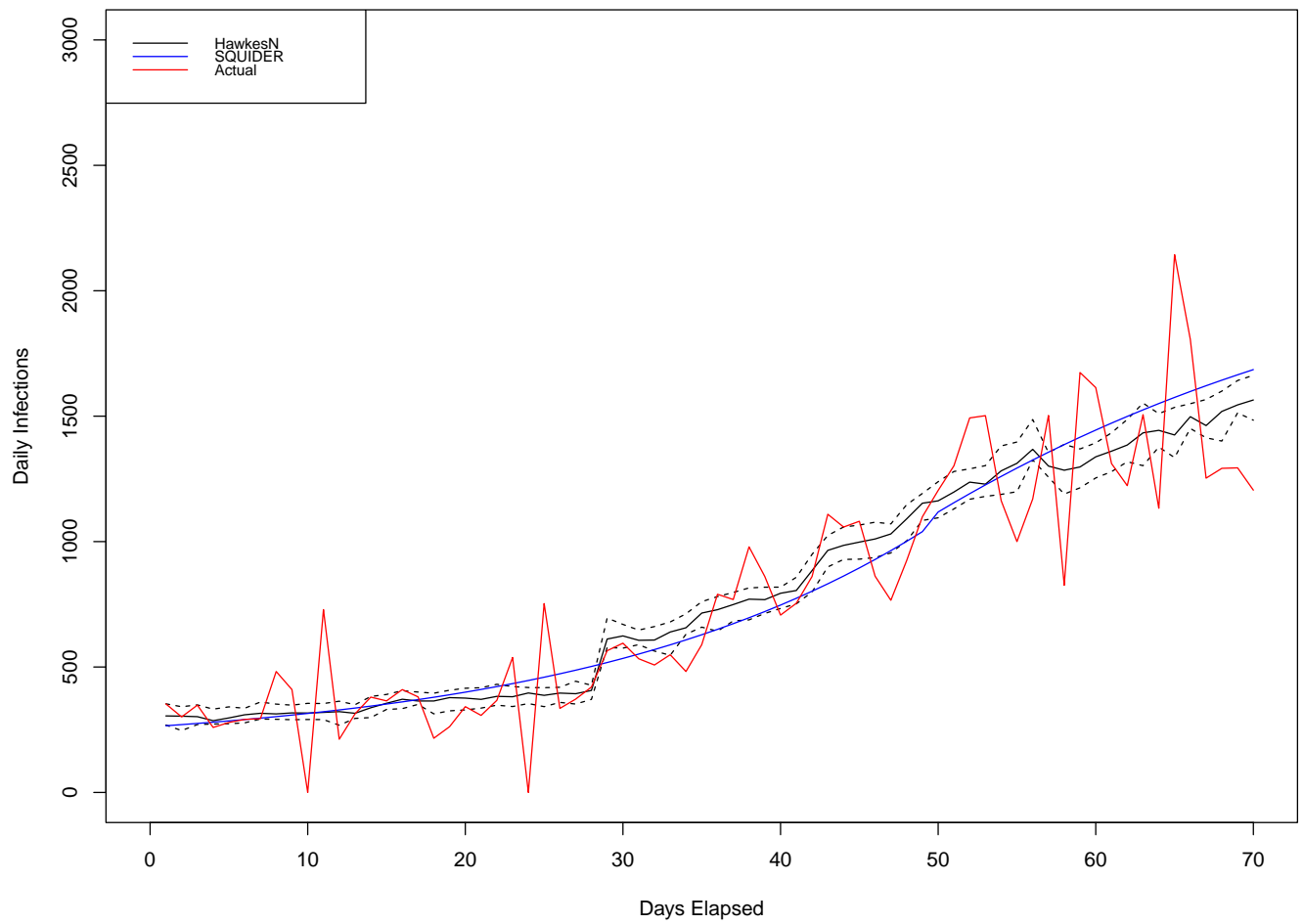
Figure 5.1: Fit of SQUIDER and Hawkes models to daily infections of COVID-19 for the first wave in Oregon.
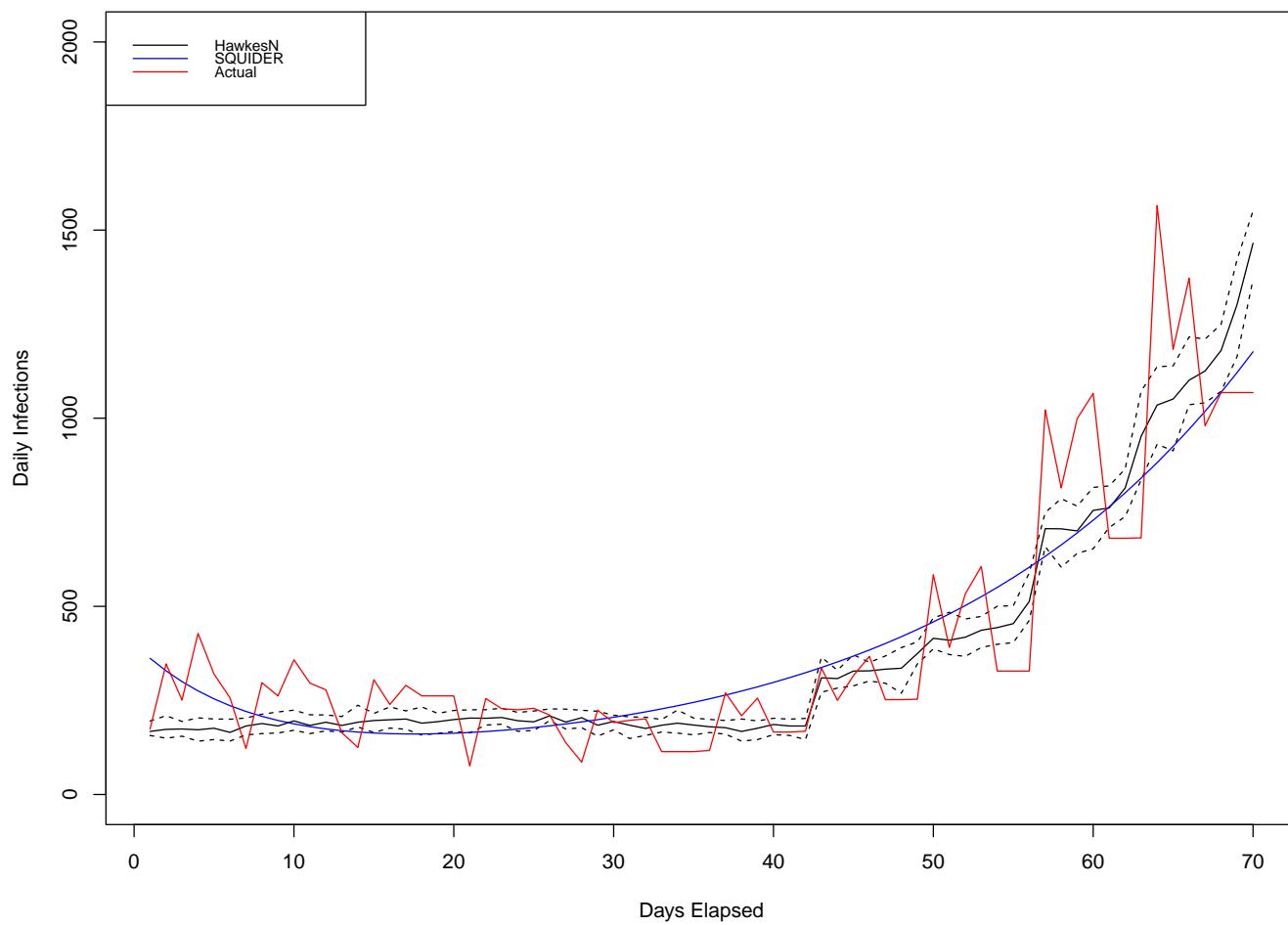
Figure 5.2: Fit of SQUIDER and Hawkes models to daily infections of COVID-19 for the second wave in Oregon.

μ	κ	g
201.553	0.469	0.032
	0.526	0.009
	1.077	0
	1.126	0.004
	0.989	0.004
		0.292
		0
		0.108
		0.112
		0.000
		0.000
		0.012
		0.038
		0.388

Table 5.1: μ, κ, and g parameters for the HawkesN model for the first wave fit of Oregon (values are rounded to 3 decimal places).

μ	κ	g
119.977	0.423	0
	0.416	0.001
	0.324	0.001
	1.097	0.039
	1.484	0.026
		0.042
		0.597
		0
		0
		0
		0.029
		0.013
		0.013
		0.238

Table 5.2: μ, κ, and g parameters for the HawkesN model for the second wave fit of Oregon (values are rounded to 3 decimal places).

Parameter	First Wave Values	Second Wave Values
a	1.001	1.004
β	0.713	0.743
δ	0.544	0.610
ϵ	0.113	0.049
γ	0.000	0.001
ρ	0.157	0.033
α	0.003	0.129
U_0	0.033	0.077
$q1$	0.006	0.024
$q2$	-0.723	-0.682
$q3$	0.094	0.011
$q4$	-0.005	-0.445
Cumulative Cases	89499.020	229654.500
Cumulative Deaths	1075.599	2951.702

Table 5.3: SQUIDER-Model Parameters for the first and second wave fits of Oregon (values are rounded to 3 decimal places).

Figure 5.3: Fit of SQUIDER and Hawkes models to daily infections of COVID-19 for the first wave in South Carolina.
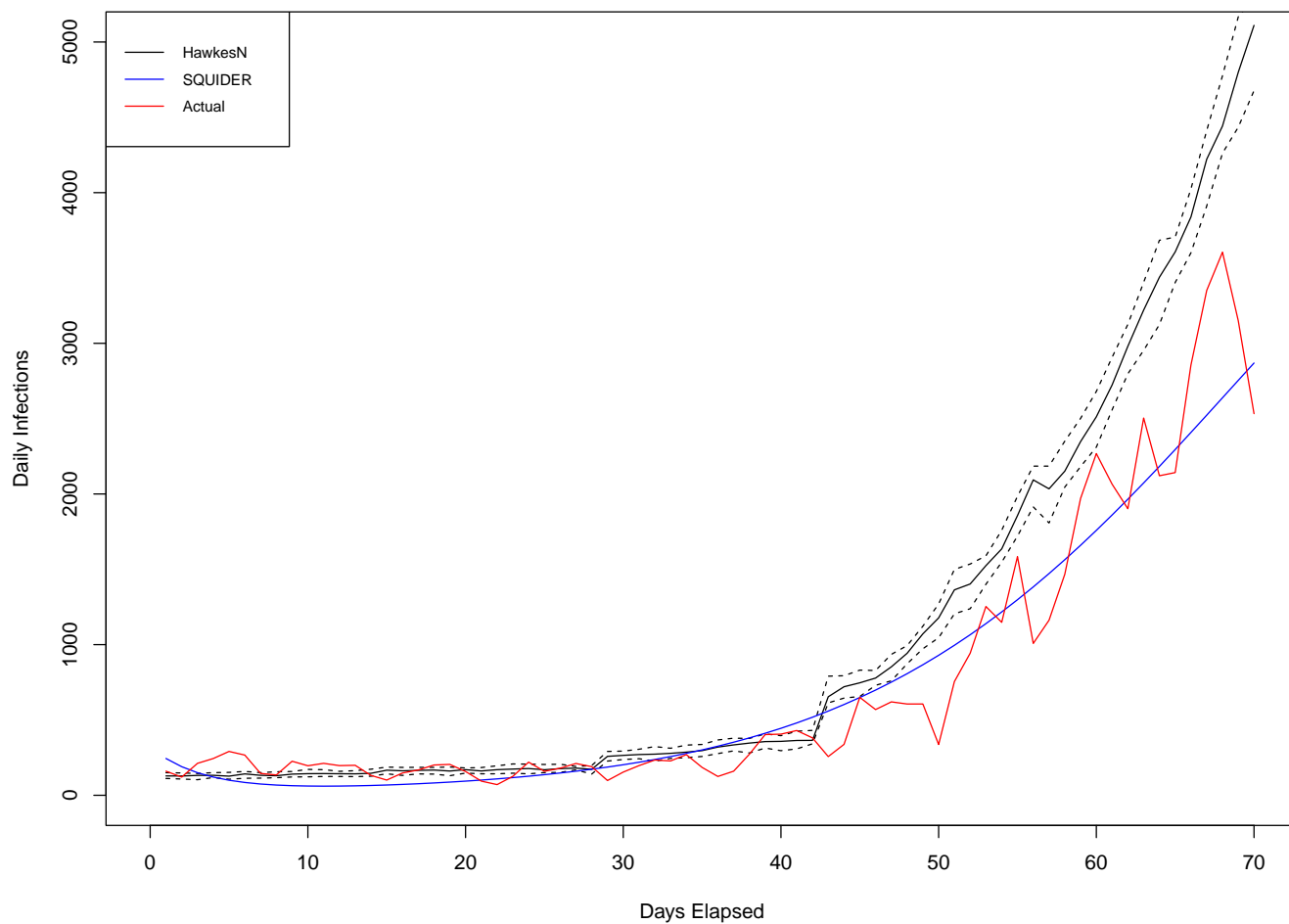
Figure 5.4: Fit of SQUIDER and Hawkes models to daily infections of COVID-19 for the second wave in South Carolina.

μ	κ	g
340.278	0.628	0.502
	0.698	0
	0.726	0.06
	0.863	0.012
	1.003	0.003
		0
		0.183
		0
		0
		0.131
		0.003
		0.105
		0.000
		0.000

Table 5.4: μ, κ, and g parameters for the HawkesN model for the first wave fit of South Carolina (values are rounded to 3 decimal places).

μ	κ	g
90.664	0.416	0.158
	0.507	0.007
	0.943	0
	1.673	0.055
	1.512	0.074
		0.142
		0.061
		0.295
		0
		0
		0.005
		0.017
		0.186
		0

Table 5.5: μ, κ, and g parameters for the HawkesN model for the second wave fit of South Carolina (values are rounded to 3 decimal places).

Parameter	First Wave Values	Second Wave Values
a	0.998	0.983
β	0.699	0.775
δ	0.525	0.520
ϵ	0.104	0.250
γ	0.001	0.001
ρ	0.004	0.041
α	0.000	0.261
U_0	0.014	0.048
$q1$	0.003	0.000
$q2$	-0.511	-0.691
$q3$	0.096	0.016
$q4$	-0.455	-0.308
Cumulative Cases	243153.500	643739.500
Cumulative Deaths	4783.332	10132.400

Table 5.6: SQUIDER-Model Parameters for the first and second wave fit of South Carolina (values are rounded to 3 decimal places).
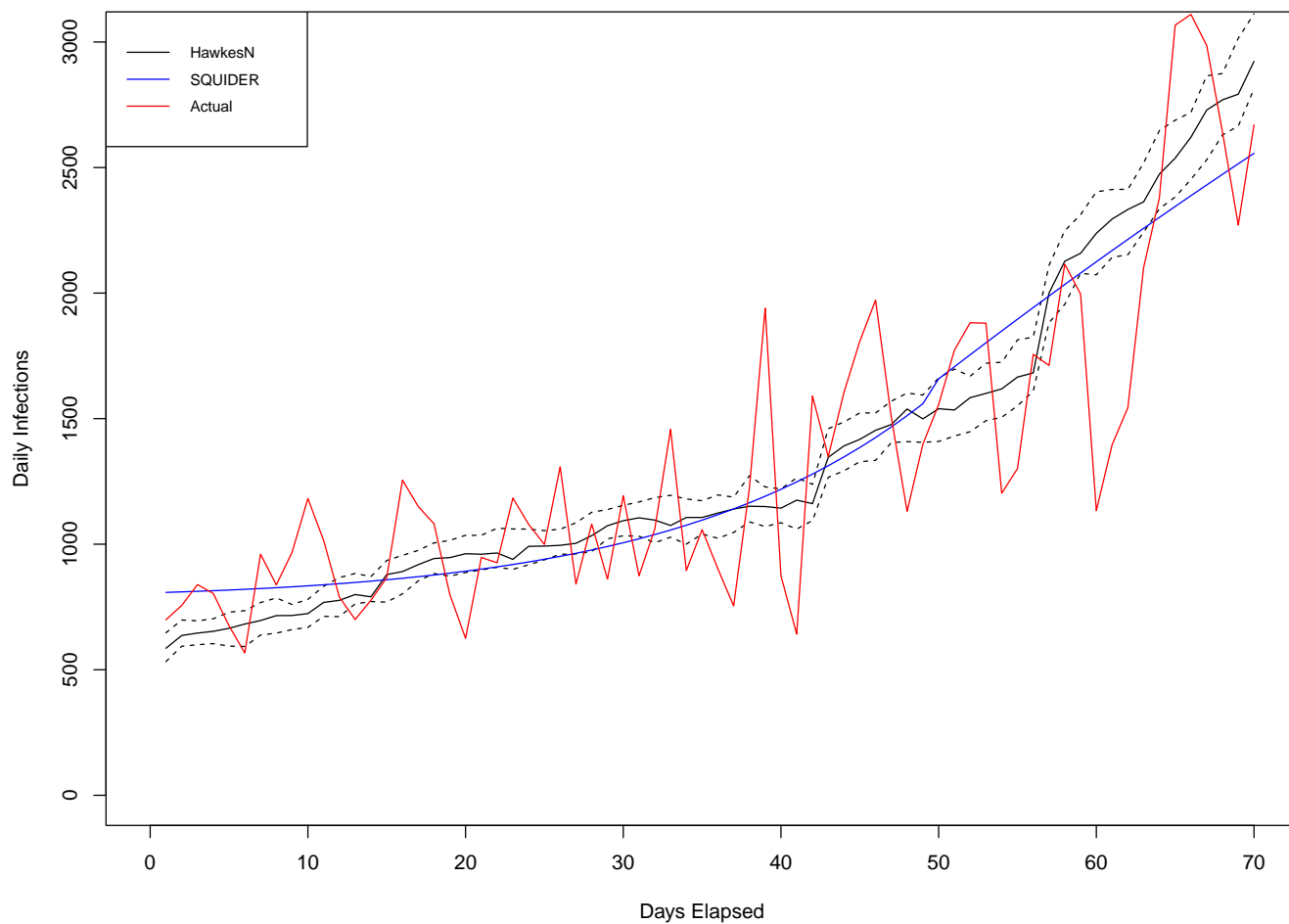
Figure 5.5: Fit of SQUIDER and Hawkes models to daily infections of COVID-19 for the first wave in Washington.
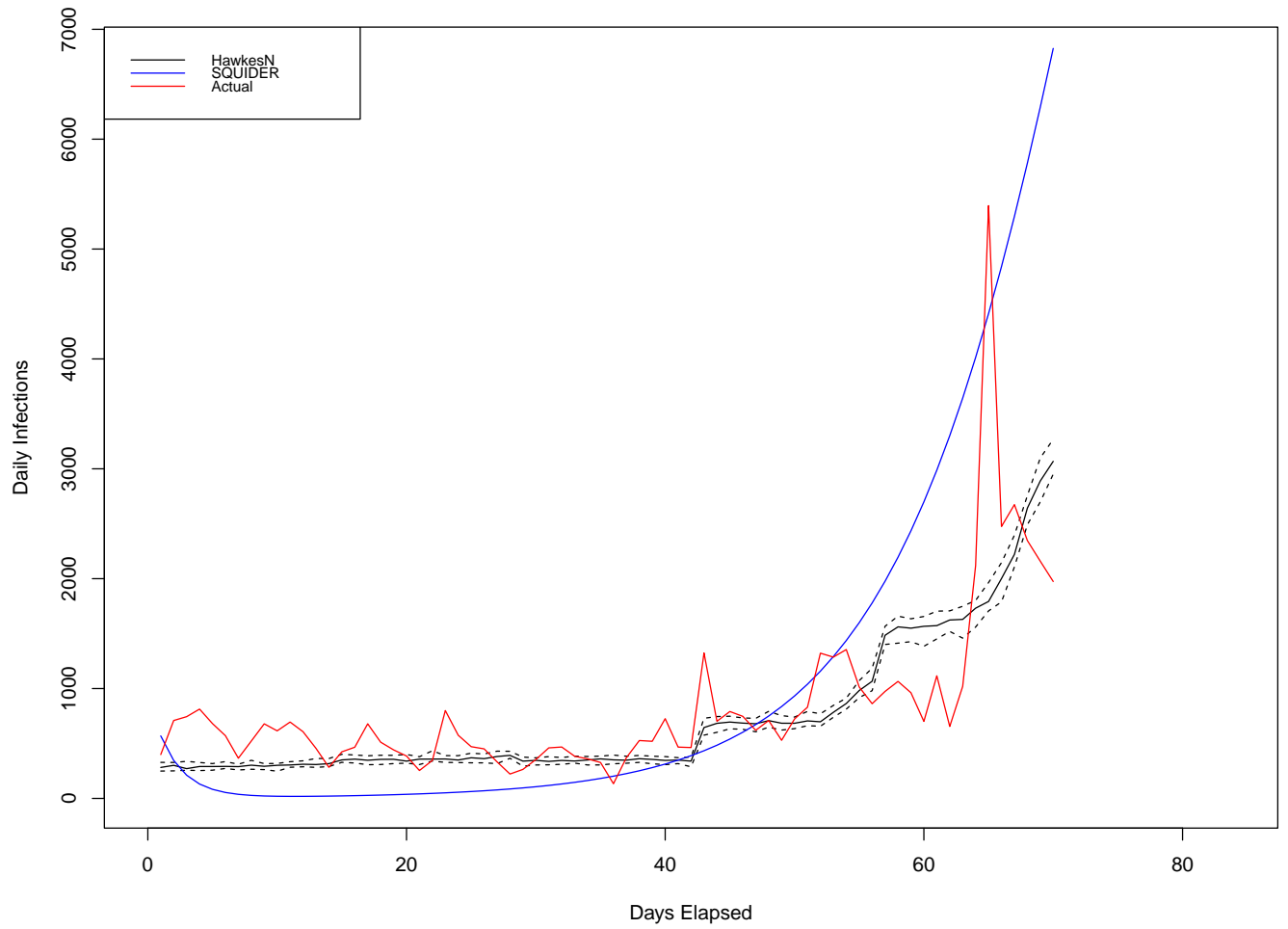
Figure 5.6: Fit of SQUIDER and Hawkes models to daily infections of COVID-19 for the second wave in Washington.

μ	κ	g
292.558	0.602	0.137
	0.673	0.06
	1.113	0
	1.387	0.003
	1.028	0
		0.004
		0.016
		0
		0.026
		0.094
		0.000
		0.039
		0.000
		0.622

Table 5.7: μ, κ, and g parameters for the HawkesN model for the first wave fit of Washington (values are rounded to 3 decimal places).

μ	κ	g
195.732	0.451	0.147
	0.515	0
	0.431	0.003
	1.229	0.003
	1.667	0.006
		0.003
		0
		0.011
		0
		0
		0.351
		0.094
		0.382
		0

Table 5.8: μ, κ, and g parameters for the HawkesN model for the first wave fit of Washington (values are rounded to 3 decimal places).

Parameter	First Wave Values	Second Wave Values
a	0.995	1.000
β	0.714	0.714
δ	0.600	0.492
ϵ	0.043	0.085
γ	0.001	0.000
ρ	0.558	0.018
α	0.692	0.397
U_0	0.091	0.010
$q1$	0.000	0.048
$q2$	-0.901	-0.491
$q3$	1.000	0.059
$q4$	-0.998	-0.495
Cumulative Cases	202179.000	511200.300
Cumulative Deaths	3132.268	6058.945

Table 5.9: SQUIDER-Model Parameters for the first and second wave fit of Washington (values are rounded to 3 decimal places).
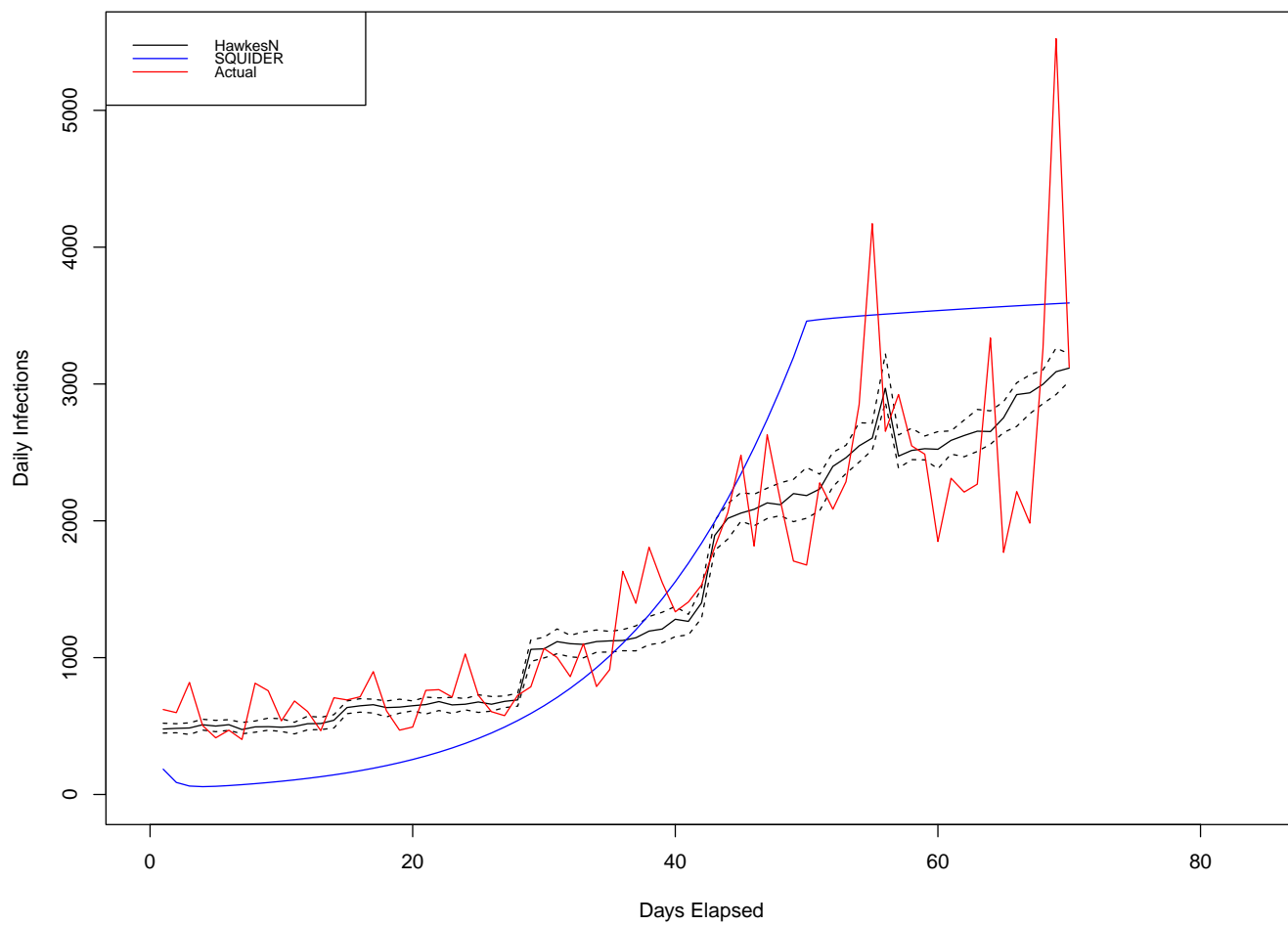
Figure 5.7: Forecast of SQUIDER and Hawkes models to daily infections of COVID-19 for the first wave over the 10 lagged days since the fitting ended in Oregon.

Figure 5.8: Forecast of SQUIDER and Hawkes models to daily infections of COVID-19 for the second wave over the 10 lagged days since the fitting ended in Oregon.
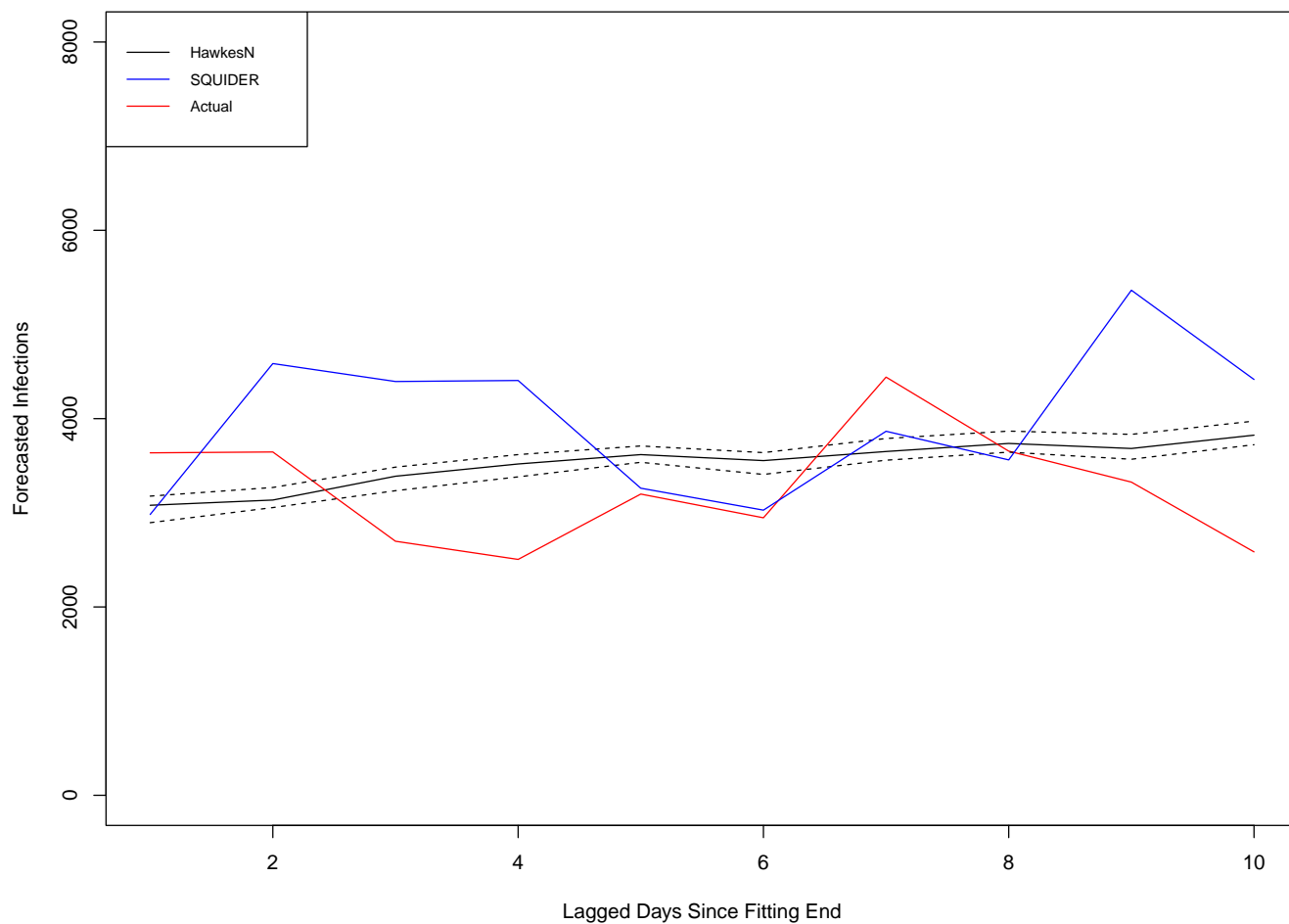
Figure 5.9: Forecast of SQUIDER and Hawkes models to daily infections of COVID-19 for the first wave over the 10 lagged days since the fitting ended in South Carolina.
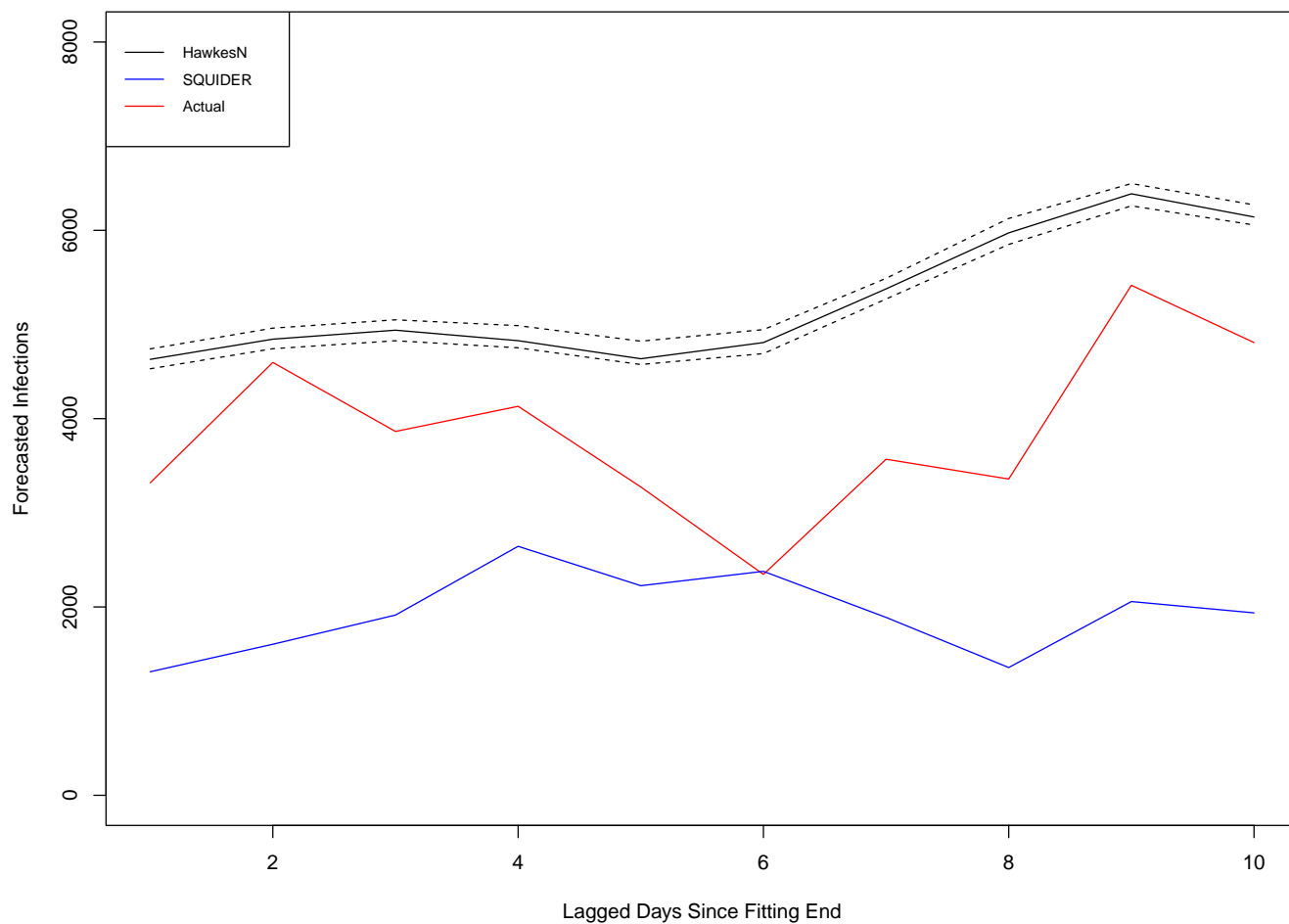
Figure 5.10: Forecast of SQUIDER and Hawkes models to daily infections of COVID-19 for the second wave over the 10 lagged days since the fitting ended in South Carolina.
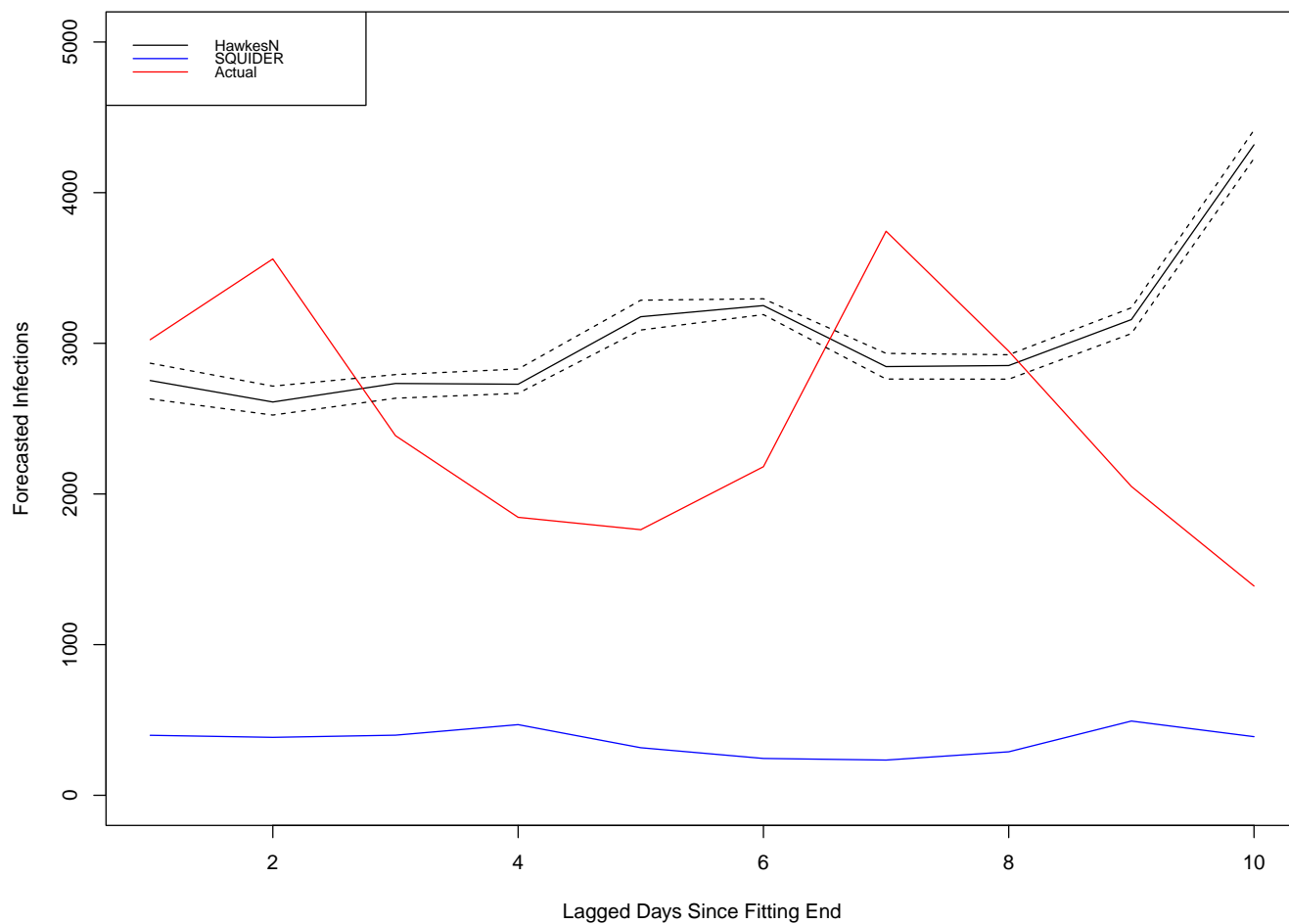
Figure 5.11: Forecast of SQUIDER and Hawkes models to daily infections of COVID-19 for the first wave over the 10 lagged days since the fitting ended in Washington.
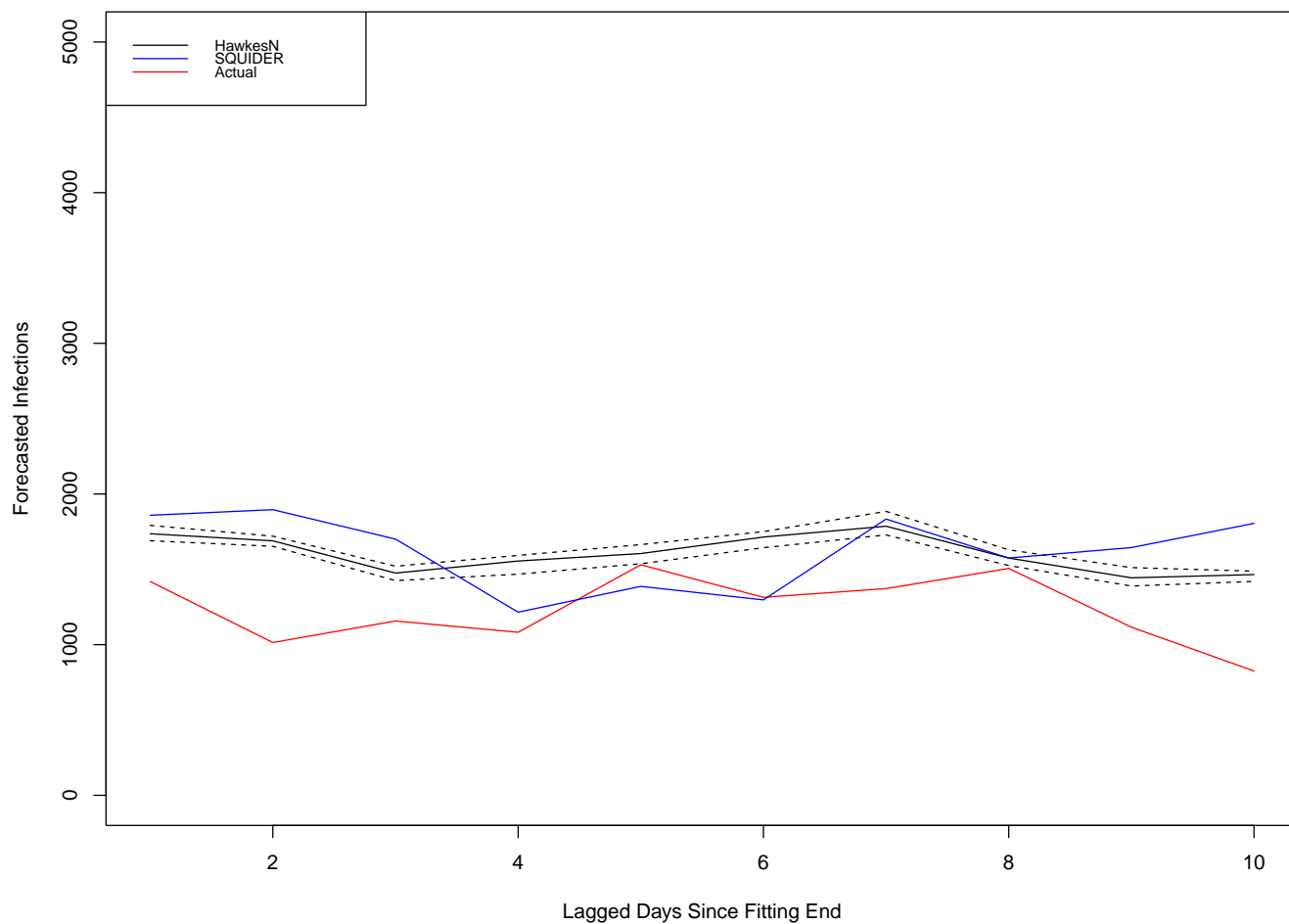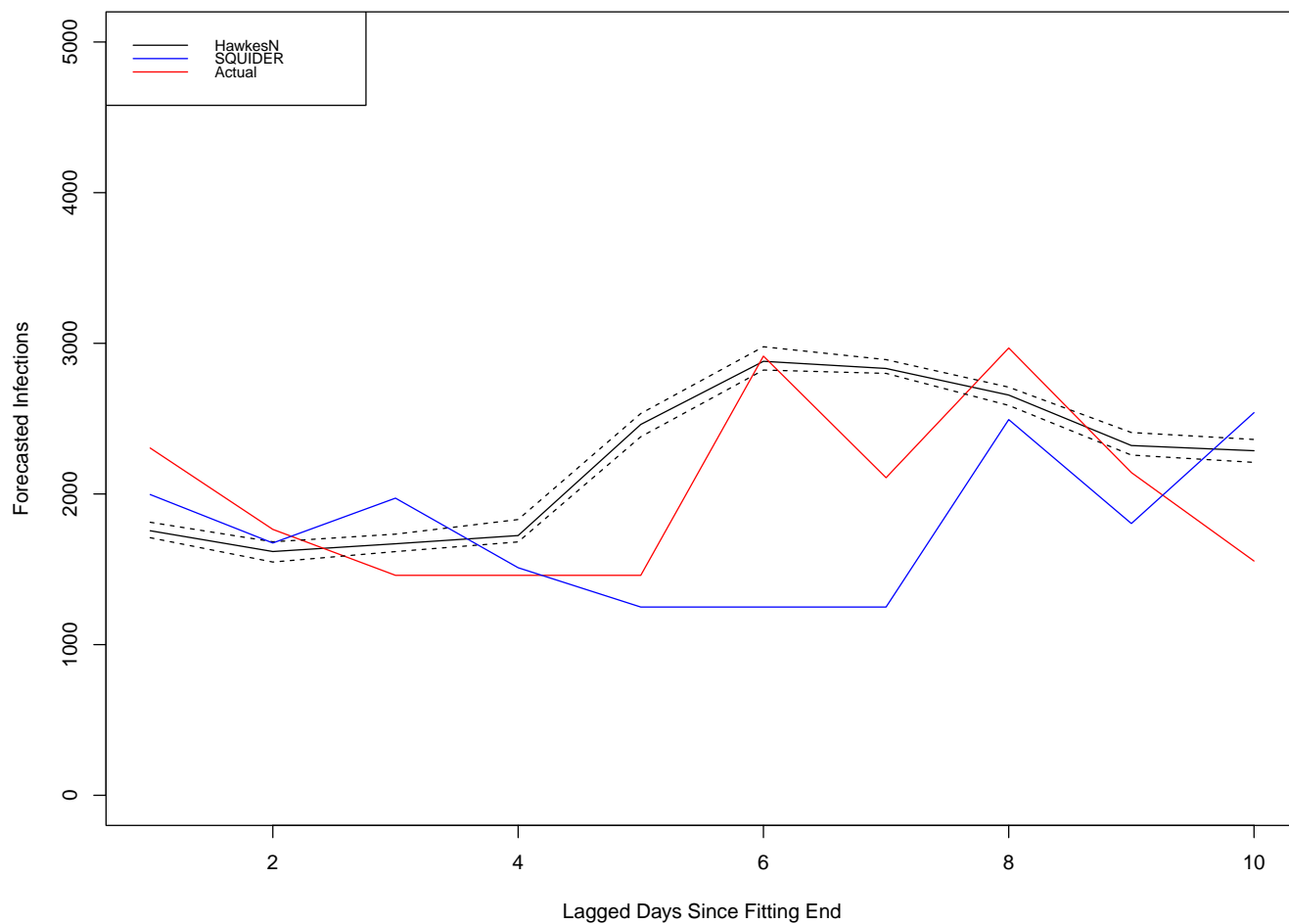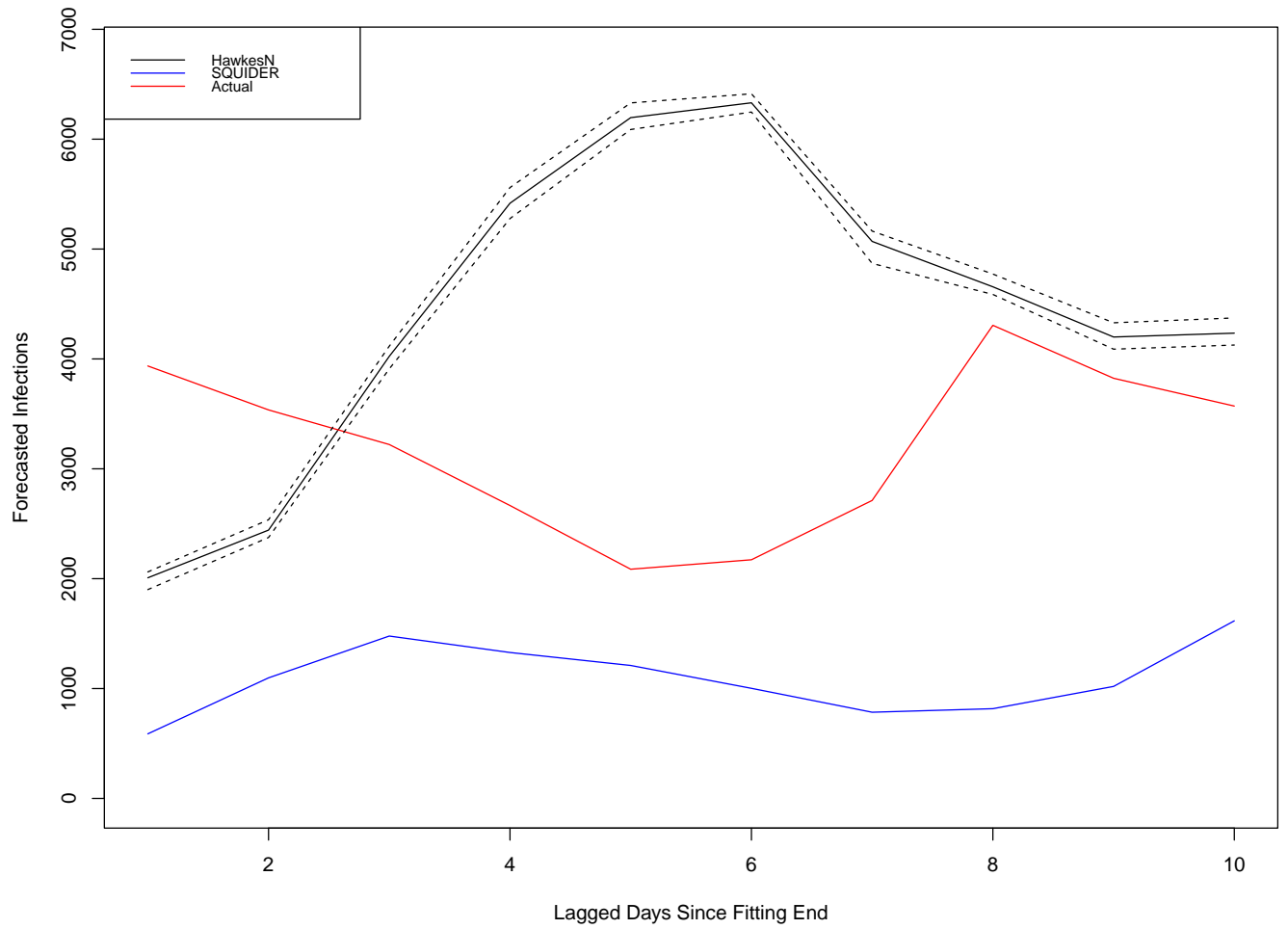
Figure 5.12: Forecast of SQUIDER and Hawkes models to daily infections of COVID-19 for the second wave over the 10 lagged days since the fitting ended in Washington.

State	Wave	Model	1st	2nd	3rd	4th	5th
Oregon	1	HawkesN	152.52	161	99.91	181.07	327.2
Oregon	1	SQUIDER	154.68	165.57	142.17	194.18	340.42
Oregon	2	HawkesN	120.74	72.61	53.12	109.64	268.9
Oregon	2	SQUIDER	90.72	81.29	89.15	177.43	419.58
South Carolina	1	HawkesN	194.61	195.6	345.54	290.25	519.19
South Carolina	1	SQUIDER	148.8	213.91	337.37	334.46	500.53
South Carolina	2	HawkesN	77.26	44.35	101.15	506.58	1171.35
South Carolina	2	SQUIDER	118.78	73.81	104.86	230.1	532.56
Washington	1	HawkesN	169.81	144.74	286.55	520.03	840.65
Washington	1	SQUIDER	525.89	427.91	319.5	898.56	1247.46
Washington	2	HawkesN	324.28	182.58	149.95	321.33	1149.425
Washington	2	SQUIDER	493.8	429.22	353.87	353.86	2664.19

Table 5.10: Fitting RMSEs of both models for all states and both waves (values are rounded to 2 decimal places).

State	Wave	HawkesN	SQUIDER
Oregon	1	417.07	525.51
Oregon	2	513.04	723.44
South Carolina	1	701	1251.19
South Carolina	2	1554.54	2155.19
Washington	1	1250.46	2269.07
Washington	2	2314.68	2273.8

Table 5.11: Forecasting RMSEs of both models for all states and both waves (values are rounded to 2 decimal places).
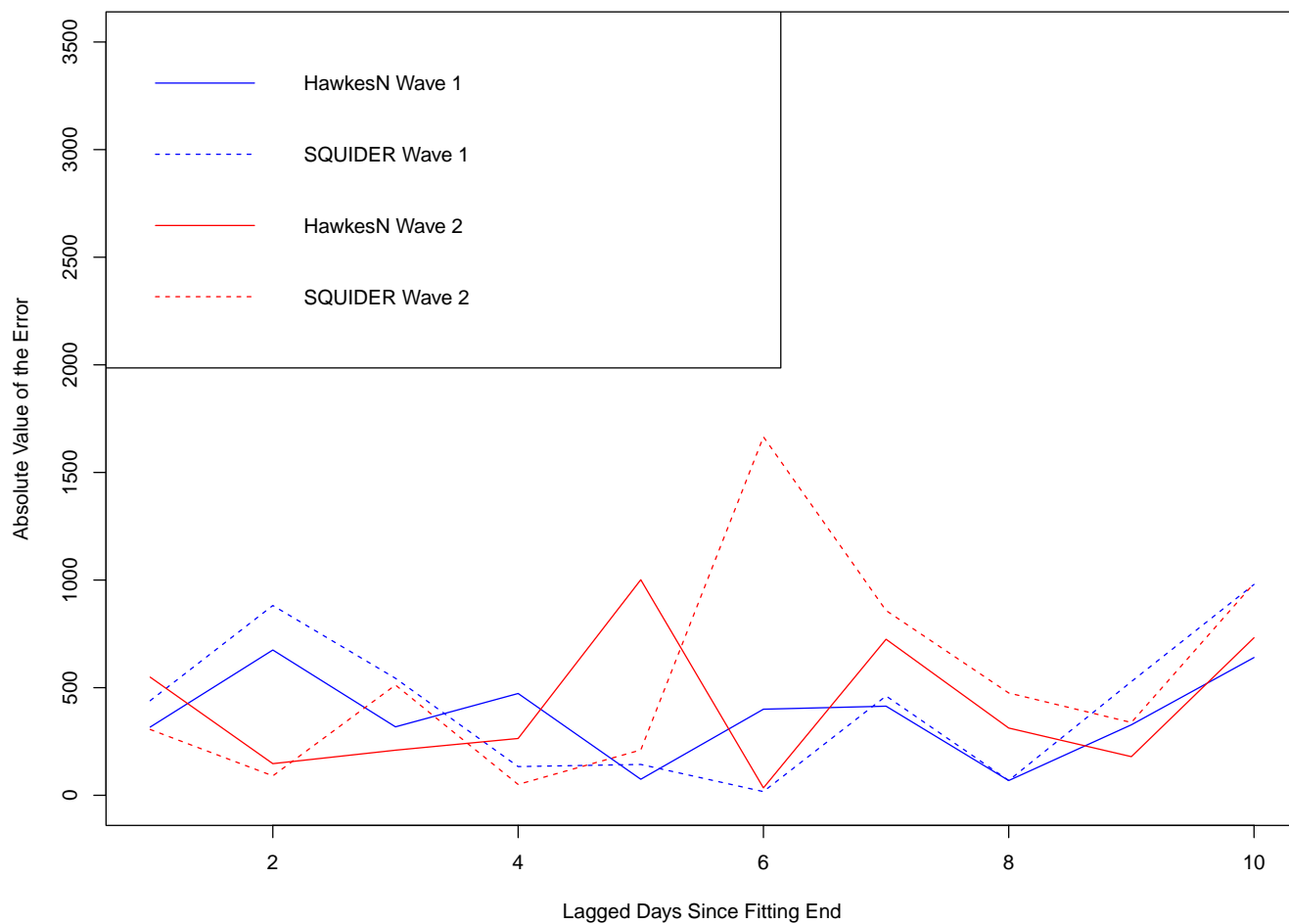
Figure 5.13: Absolute error values for SQUIDER and HawkesN during both waves of COVID-19 in Oregon.
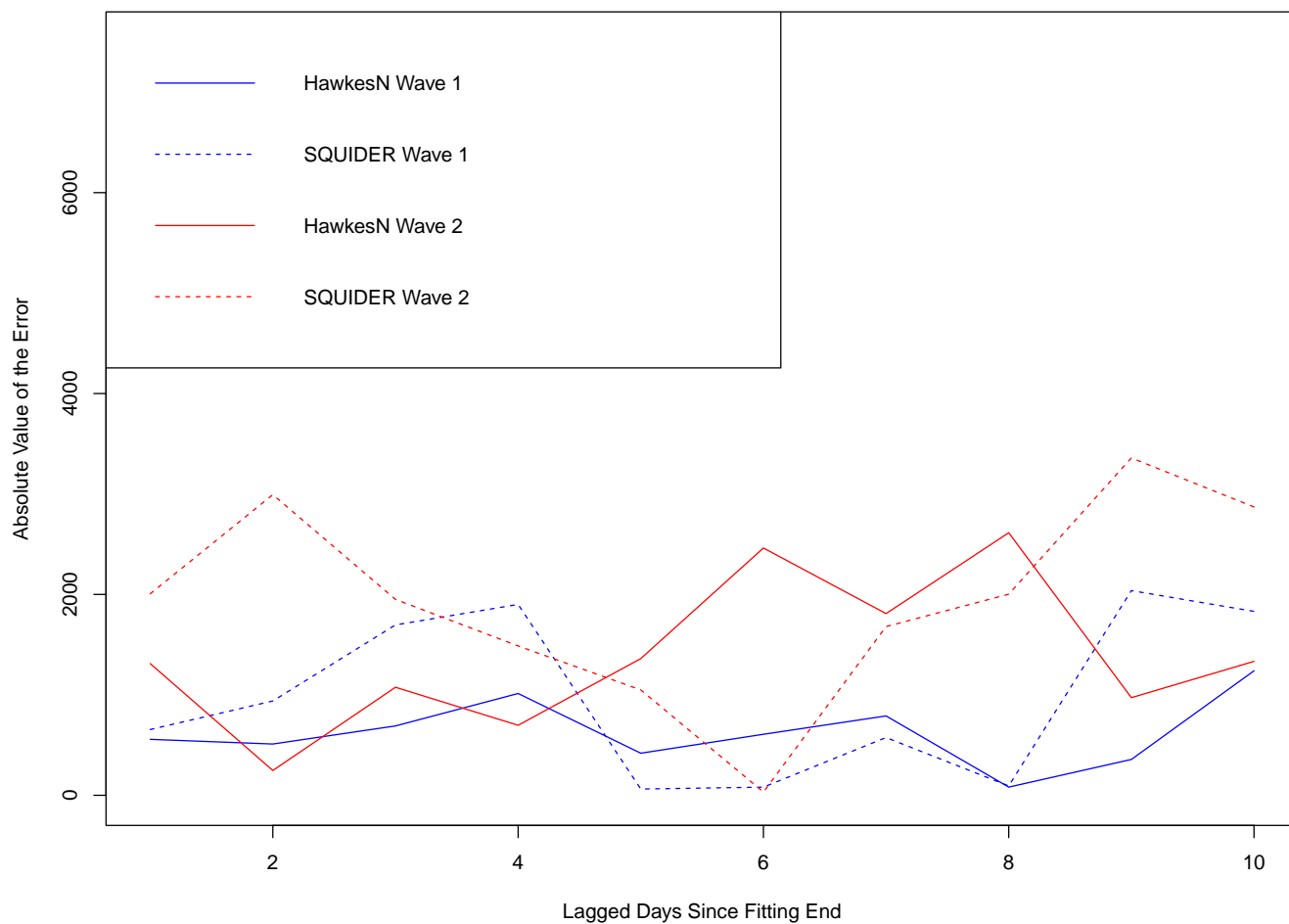
Figure 5.14: Absolute error values for SQUIDER and HawkesN during both waves of COVID-19 in South Carolina.
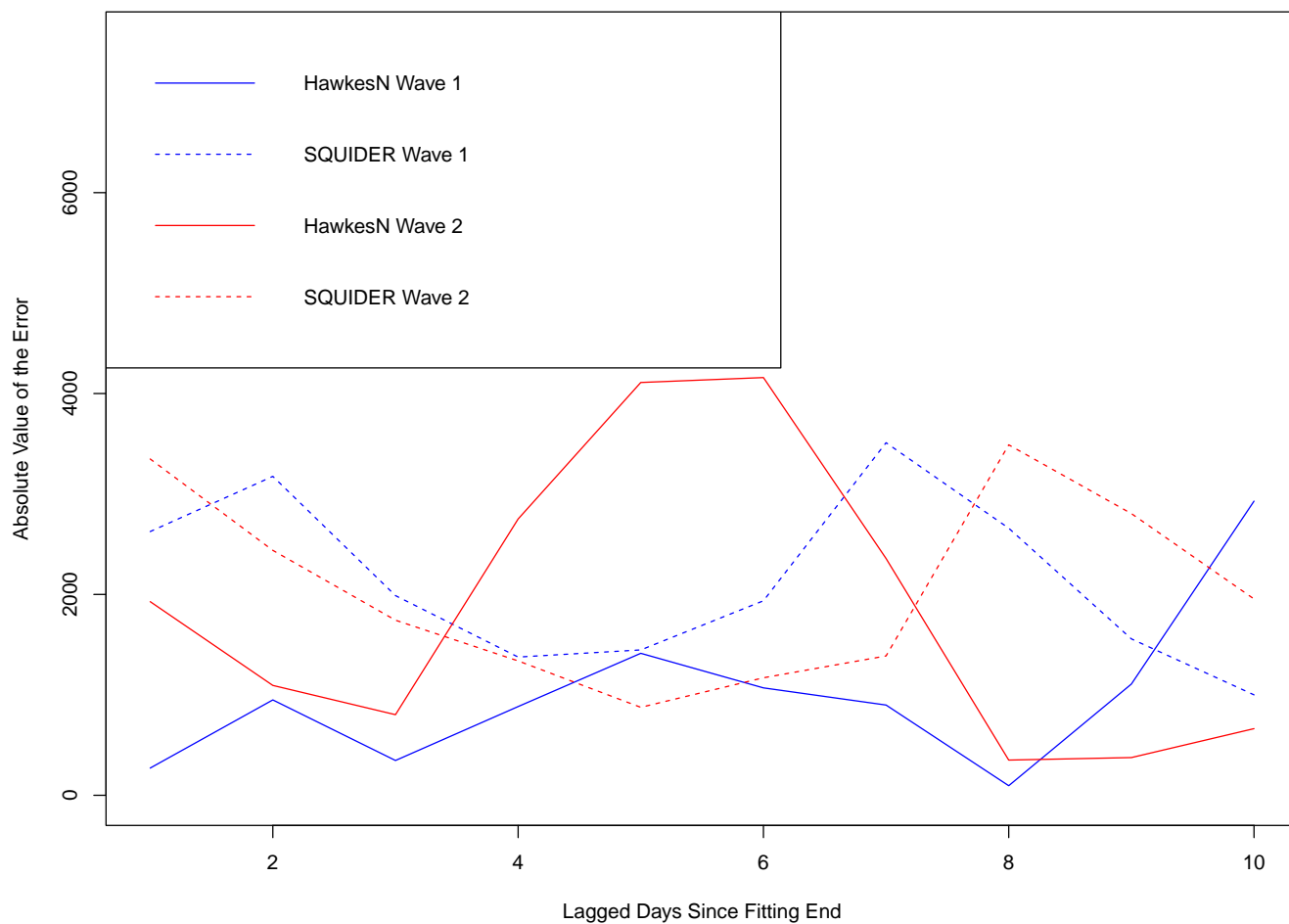
Figure 5.15: Absolute error values for SQUIDER and HawkesN during both waves of COVID-19 in Washington.

5.1 Fitting

It is important to note the dates in which the first and second waves took place. We say for the purposes of this thesis, that the first wave of COVID-19 started on the 1st October 2020. The second wave started on the 1st June 2021, where we take a look at the delta variant. As seen in

the legends of the above plots, we have that the actual daily infections is indicated by the red line, the SQUIDER model is indicated by the blue line and the HawkesN model is indicated by the black line, with the dashed black lines indicating the 5% and the 95% lower and upper bound values.

5.1.1   Oregon First Wave

We see that the first wave of COVID-19 in Oregon was fit very well by both the HawkesN and SQUIDER Model, as indicated by figure 5.1. We see that whilst the overall trend was followed very well, that that the individual peaks of daily cases wasn't followed very well. This is especially seen for the SQUIDER model as there is no fluctuations in the plot, rather, one continuous line. This is the result of fitting twelve parameters, using the *optim* function in R, to noisy data, such as, the daily case counts of COVID-19. The fluctuations seen in the black lines show that whilst it follows the general trend of the peaks and troughs, it is no where near as extreme as the actual values. We see that for all 5 of the 14-day fitting periods, SQUIDER performs slightly worse than the HawkesN model as indicated by the slightly higher values of RMSE. We notice that between 60 and 70 days elapsed, there is more variance in the number of actual daily infections, however whilst this is the case, both HawkesN and SQUIDER seem to fit the daily case counts fairly well.

5.1.2   Oregon Second Wave

We see that the second wave of COVID-19 in Oregon was fit very well by both the HawkesN and SQUIDER Model, as indicated by figure 5.2. We see that whilst the overall trend was followed very well, that that the individual peaks of daily cases wasn't followed very well. This is especially seen for the SQUIDER model as there is no fluctuations in the plot, rather is one continuous line. The fluctuations seen in the black lines show that whilst it follows the general trend of the peaks and troughs, it is no where near as extreme as the actual values. We see that at around 65 days elapsed and further, that the HawkesN model does not seem to take into account the downward

trend seen by the actual, that is, the red line, but instead carries on trending upwards. We see that, with the exception of the first of the 14-day fitting periods, that 4 of the 14-day fitting periods, SQUIDER performs slightly worse than the HawkesN model as indicated by the slightly higher values of RMSE. We can see the slight improvement in the RMSE of 90.72 for SQUIDER compared to that of 120.74 from HawkesN, since the actual overall trend is negative for the first 14-days, which is also seen by SQUIDER's initial negative gradient. The variation of actual daily case counts of COVID-19 was less than that of the first-wave for Oregon. This indicates that the shape of the curve fit matters more.

5.1.3 South Carolina First Wave

The fit for the first wave of South Carolina, from figure 5.3, is seen to be pretty good. Overall we see that the plots for HawkesN follows the overall pattern of the actual number of cases very closely, except for the circa 60 days elapsed mark, where the trough in the number of cases, is not mimicked at all by the HawkesN model, but instead continues to rise. This lack of accurate fitting, is then corroborated in the RMSE values, where for the fifth fitting period, we see a higher RMSE value compared to that of the SQUIDER model (519.19, compared to 500.53). It is also important to note, that, although the SQUIDER model does not follow the peaks and troughs of the daily infection counts, it does a very good job of following the smoothed out pattern. During the fifth 14-day fitting period, we see a larger variation in the estimated daily infections from the 20 simulations. This is indicated by the larger gap between the two black dotted lines on figure 5.3, compared to the rest of the plot.

5.1.4 South Carolina Second Wave

From figure 5.4, we had an overall great fit for the second wave of COVID-19 in South Carolina, indicated by the very close following of the lines for both HawkesN and SQUIDER models to the actual number of daily infections. As seen in table 5.10, up to the fourth 14-day fitting period, we

43

see that the HawkesN model outperforms the SQUIDER model, however from the fourth fitting period onwards, SQUIDER performs better in fitting the number of daily infections of COVID-19. This is represented graphically due to the steeper positive gradient seen by the HawkesN line(s) compared to that of the actual number of cases line. This is likely due to an overcompensation in the productivity, κ, coefficient, when the slope of the case counts increased, as seen in table 5.5. Note, that g is fit to all 5 14-day fitting periods simultaneously, but κ is fit for each individual fitting period, thus making this aforementioned situation, plausible.

5.1.5 Washington First Wave

In fitting Washington's first wave of daily infections, taking a look at figure 5.5, that we found that across all fitting periods, HawkesN performed better than the SQUIDER model, indicated by the lesser RMSE values in table 5.10. This is especially evident at the beginning and at the end of our overall fitting period, where the blue line SQUIDER is a lot farther apart from he red line indicating the actual cases. Taking a look more at the end of the fit, we see that the very big peak in actual case numbers is not captured very well by the HawkesN model. The shape of the blue line, is determined by the fact that there were a lot more people going into quarantine during the fall and winter surge of 2020. One disadvantage of fitting the SQUIDER model in this scenario, is that curve it estimates before the quarantine period is overestimated by the surge that is seen in the number of daily infections from roughly 30 days elapsed and onwards.

5.1.6 Washington Second Wave

Taking a look at figure 5.6, we find that once again, we have found that HawkesN performs considerably better than the SQUIDER model at all points in the overall fitting period, with the exception of the last peak indicated on plot. HawkesN seem to follow the overarching trend more closely, but the SQUIDER model does a better job at fitting the sudden peak in the actual number of cases. We find that the number of COVID-19 cases given by the SQUIDER and the

actual case counts are equal at roughly 65 days elapsed. Due to a low concavity, a coefficient, we also notice that the SQUIDER fit decreases during the first 5-10 days of the total fitting period, likely overcompensating for the peak at circa 65 days elapsed.

5.2 Forecasting

5.2.1 Oregon First Wave

From figure 5.7, we see that the first wave of COVID-19 in Oregon was forecasted fairly well. As expected, HawkesN performed better with regards to RMSE with a value of 417.07, compared to SQUIDER's 525.51. Whilst this is the case for the overall forecasting RMSE, we see that between the 4 and 6 lagged days points, it seems that the SQUIDER model performs slightly better than HawkesN. Both HawkesN and SQUIDER overestimate the number of daily cases of COVID-19 - this is because the in the last fitting period for the first wave in Oregon, we have a steeper gradient than that of the forecasting period.

5.2.2 Oregon Second Wave

We see that the second wave of COVID-19 in Oregon was forecasted worse than the first wave, as indicated in figure 5.8. This is, however, to be expected, as there were a larger number of cases of COVID-19. We see an approximate 2x larger RMSE value for SQUIDER compared to that of HawkesN. This is most likely, as seen in figure 5.8, due to the 4 lagged days point on the graph until the 7 lagged days point on the graph, where there is almost no following of the overall trend of the actual forecast. Whilst we have that the HawkesN model overestimates the number of daily infections, between 4 and 5 lagged days, the peaks at 6 and 8 lagged days since the Fitting End, are forecasted correctly.

5.2.3 South Carolina First Wave

Forecasting the first wave of COVID-19 in South Carolina, turned out overall, pretty successful, at least for the HawkesN model as indicated by figure 5.9. We see from table 5.10, that SQUIDER produces an RMSE value of roughly double, 1251.19, that of the HawkesN model, 701, indicating a worse forecasting. Saying that, there was something around the fifth and sixth lagged day since the fitting ended, that allowed the SQUIDER to far outperform the HawkesN model, as seen by the very small distance between the red and blue lines during those aforementioned points. Other than from between 5 lagged days and 8 lagged days since the fitting end, the SQUIDER model overestimates the actual daily case counts of COVID-19, whereas the HawkesN model forecasts closer to that of the actual cases.

5.2.4 South Carolina Second Wave

From table 5.11, SQUIDER performed worse in forecasting the daily case counts of COVID-19 for the second wave in South Carolina. We see however, that whilst overall SQUIDER performs worse, taking a look at the sixth lagged day since the fitting end, that the forecasted infections given by the SQUIDER model are exactly the same as the actual forecasted infections. During the final fitting period, the SQUIDER model emulates the actual case counts of COVID-19 fairly accurately, whereas the HawkesN model overestimates the number of daily case counts. However, during the forecasting period, the actual case count increases at a faster rate than during the final 14-day fitting period. This allows for the HawkesN forecast to outperform the SQUIDER forecast, despite the better fit from the SQUIDER model in the last fitting period.

5.2.5 Washington First Wave

Taking a look at figure 5.11, we see that the SQUIDER model performed extremely poorly in forecasting the number of forecasted infections, which is confirmed by the extremely high RMSE value for forecasting the Washington first wave out our forecasting table 5.11. Saying this, the

HawkesN model does not do great either, with only between 5 and 6 lagged days, does the graph follow even the same trend. However due to the fact that the error at each point between the actual forecasted infections and the HawkesN forecasted infections is quite obviously less than between actual and SQUIDER, we end up concluding that HawkesN outperforms SQUIDER in forecasting the first wave of COVID-19 in Washington State.

5.2.6 Washington Second Wave

In forecasting the second wave of COVID-19 daily infection case counts in Washington, just looking at the figure 5.12, we saw, in this case, that the SQUIDER model actually outperforms the HawkesN model. We can see that the HawkesN model only starts to forecast in a similar trend to the actual cases at the eighth lagged day after the fitting ended. Whilst from this point on there was a very good forecast, the amount of discrepancy in even forecasting the same trend of daily infections, meant that the RMSE increase above SQUIDER's who at least for the most part forecasted the same trends. This is confirmed by looking at table 5.11, where we see an RMSE of 2314.68 for HawkesN compared to SQUIDER's 2273.8.

5.3 Absolute Values of the Errors

Figures 5.13, 5.14, and 5.15, display the absolute values of the errors between the actual and forecasted case counts of COVID-19. Each figure displays the absolute values of the errors of both models for both the first and second wave, with as the legend shows, the first wave data being represented by the blue colored lines, and the second wave data being represented by the red colored lines.

5.3.1 Oregon

Due to the number of daily infections in the first wave being a lot smaller than the second wave, as expected we see from figure 5.13, that the absolute values of the errors at each lagged day since the fitting end is much smaller than the second wave. For the first wave, we see that our minimum is at the 6 and 8 lagged days since fitting end point with an error absolute value of nearly 0 for both the SQUIDER model and HawkesN model respectively. The maximum is seen to be reached at 10 and 2 lagged days for the SQUIDER model and the HawkesN model respectively, with the SQUIDER model giving a maximum absolute value of error of roughly 1000, whereas the for the HawkesN model, an absolute value of error of roughly 600. With the exception of the fourth and sixth lagged days since the fitting end, we see that the HawkesN model performs better than the SQUIDER model. For our second wave of COVID-19 in Oregon, we see that a minimum value of absolute error is reached at 4 lagged days, and 6 lagged days, since the fitting end, for the SQUIDER model and the HawkesN model respectively. On the other hand, our maximum absolute values of error are found to be at 6 lagged days, and 5 lagged days for the SQUIDER model and HawkesN model respectively. We see that, with the exception of the second, fourth and fifth lagged day since the fitting end, that HawkesN performs better in regards to forecasting. The excellence of HawkesN's forecasting capabilities, in comparison to SQUIDER's capability should be duly noted.

5.3.2 South Carolina

With the exception of the the SQUIDER model at the fourth lagged day since the fitting end, we see that overall, the forecasts of daily case counts of COVID-19 for the first wave in South Carolina, are seen to perform better, as indicated by the lower absolute values of error, compared to the second wave. For our first wave, we see that our minimum is at the 5 and 8 lagged days since the fitting end, giving absolute values of error of roughly 0 for both models. Maximums for the first wave are attained at the 9 and 10 lagged day mark, for SQUIDER and HawkesN

respectively giving absolute values of error of roughly 2000 and 1500. Moving on to the second wave, we see that the minimum absolute value of errors are attained for the SQUIDER model, at the 6 lagged day since fitting end mark, giving a value of roughly zero, which is similar to the minimum absolute value of error given by the HawkesN model attained at 2 lagged days since the fitting end, also giving a minimum absolute value of error of roughly zero, also. Looking at the maximums, we see the SQUIDER achieves this at 9 lagged days, with a maximum absolute value of error of roughly 4000, whereas the maximum absolute value of error attained by the HawkesN is at the 8 lagged days since fitting end mark, giving a value of roughly 3000. Overall, With the exception of 5 through 8 lagged days since the fitting end, that HawkesN performs better with a lower absolute value of error compared to the SQUIDER model for both the first and second waves.

5.3.3   Washington

Since the number of daily infections in the first wave is a lot smaller than in the second wave, we see that in figure 5.14, that the absolute values of the errors at each lagged day since the fitting end is much smaller than the second wave. Taking a look specifically now at the first we, it is clear to see that our minimum is at the 4 and 8 lagged days since fitting end points with an absolute error value of nearly 2000 and 0 for the SQUIDER and HawkesN models respectively. The maximum is at 7 lagged days for the SQUIDER model and 10 lagged days since the fitting end for the HawkesN model, with the absolute values of error being roughly 4000 for both models. Moving now on to our second wave of COVID-19 in Washington, we see that the minimum absolute values of error are attained at the fifth and eighth lagged day points since the fitting end, giving values of 1000 and 500, for the SQUIDER model and the HawkesN model respectively. Looking now at the maximum values of absolute value error, we now see that we have them at the 8 and 6 lagged days since the fitting end, for SQUIDER and HawkesN respectively, giving the values of again, roughly 4000 for both the SQUIDER model and the HawkesN model. We see that, compared to the first wave, that it is rather inconclusive as to which model performs

49

better, since we can see sections of the figure, where SQUIDER has a lower absolute value of error than HawkesN, and other parts in which the HawkesN model has a lower absolute value of error than SQUIDER.

CHAPTER 6

Conclusion

To conclude this thesis, we talk about the overarching results found. As expected in almost all fitting periods, and forecasting, the SQUIDER model was outperformed by the HawkesN model. This matches what we were initially trying to prove.

There were exceptions to this overall trend, such as, South Carolina during the second wave during the fifth fitting period, with SQUIDER giving an RMSE value of 532.56 compared to HawkesN's 1171.35. Another example of where SQUIDER performed slightly better was during the forecasting of Washington's second wave of COVID-19, with a value of 2273.80, compared to that of HawkesN's, 2314.68.

Each of the models that we used in this thesis had their fair share of limitations.

Some examples of SQUIDER's limitations include the following:

- Used for COVID-19 only through May 2020, whereas, at the time of submission of this thesis, we are in 2022 - this means that there could have been models in more recent times that performed better than SQUIDER, by using, for example, different or additional nodes to the SIR-family model;

- Compartmental models which have been successful in the past, have been modeled once the entire architecture of what they are modeling has been understood. In the case of this SQUIDER model, it is only now, in 2022, in which we know even the basics of COVID-19.

Some of the limitations caused by the HawkesN model are as follows:

- The least squares algorithm was fantastic at fitting the HawkesN model, with a notable exception of the second wave of COVID-19 in South Carolina, in the 5th 14-day fitting period. In this case the κ parameter was larger than the expected value were to be;

- On occasion, when the least squares algorithm fits the HawkesN model well, the optimal value for κ is < 1, when, in reality, the reproduction factor is actually > 1. This is likely due to the optimization algorithm having inital value points.

In future research, some interesting points to consider would be:

- Compare the RMSE values for different length forecasts and create a plot to reflect the discrepancies found between the different length forecasts;

- Discover an optimal design to use the least squares algorithm more effectively, where the value of the κ, is more intepretable.

Overall, with proper knowledge of the item being modeled, we see that compartmental models, can actually be useful. However, in the case of this thesis, that is, in fitting and forecasting COVID-19, we have found an alternative, and more effective model in the HawkesN model.

CHAPTER 7

Acknowledgements

Thank you to Andrew Kaplan, who collaborated on this project, as well as Conor Kresin, who formalized the algorithm used in this thesis for modeling with HawkesN.

REFERENCES

[1] Kowshik chilamkurthy. Understanding point processes. `https://kowshikchilamkurthy.medium.com/understanding-point-processes-6e3d2f6c5480`, 2020. [Online; accessed 12-April-2022].

[2] Paul Embrechts, Thomas Liniger, and Lu Lin. Multivariate hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378, 2011.

[3] J. Engelstädter. Continuous-time models in several variables. `https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population`, 2022. [Online; accessed 12-April-2022].

[4] Centers for Disease Control and Prevention. Basics. `https://www.cdc.gov/coronavirus/2019-ncov/faq.html#Basics`, 2022. [Online; accessed 12-April-2022].

[5] Centers for Disease Control and Prevention. Covid data tracker. `https://covid.cdc.gov/covid-data-tracker/#trends_dailycases`, 2022. [Online; accessed 12-April-2022].

[6] Stochastic Geometry for Wireless Applications. Lecture 2: Introduction to point processes, poisson point processes.

[7] A. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[8] A. Kaplan, C. Kresin, and F.P Schoenberg. Estimation of doubling time for sars-cov-2 in california using the hawkesn model. (in press).

[9] Z. S. Khan, F. Van Bussel, and F. Hussain. A predictive model for covid-19 spread – with application to eight us states and how to end the pandemic. *Epidemiology and Infection*, 148:e249, 2020.

[10] M. Kirchner. Hawkes and inar(∞) processes. *Stochastic Processes and their Applications*, 26(8):2494–2525, 2016.

[11] M. Kirchner. An estimation procedure for the hawkes process. *Quant Financ.*, 17(4):571–595, 2017.

[12] C. Kresin, F.P. Schoenberg, and G. Mohler. Comparison of the hawkes and seir models for the spread of covid-19. *Advances and Applications in Statistics*, 2021.

[13] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association.*, 83(401):9–27, 1988.

[14] World Health Organization. Coronavirus disease (covid-19): Masks. `https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-masks`, 2022. [Online; accessed 12-April-2022].

[15] J. Park, A. Chaffee, R. Harrigan, and F.P. Schoenberg. A non-parametric hawkes model of the spread of ebola in west africa. *J. Appl. Stat.*, pages 1–27, 2020.

[16] M.A. Rizoiu, S. Mishra, Q. Kong, M. Carman, and L. Xie. 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, 2018.

[17] F.P. Schoenberg. Introduction to point processes.

[18] F.P. Schoenberg. Estimating covid-19 transmission time using hawkes point processes. 2020.

[19] F.P. Schoenberg. Estimating covid-19 transmission time using hawkes point processes. *Annals of Applied Statistics, sub 9/21*, 2021.

[20] D. Smith and L. More. The sir model for spread of disease - the differential equation model. `https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model`, 2004. [Online; accessed 12-April-2022].

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

[22] Wikipedia. List of u.s. states and territories by population. `https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population`, 2022. [Online; accessed 12-April-2022].