Medical Research Archives

European Society of Medicine

OPEN ACCESS

Published: March 31, 2024

Citation: Kaplan A, Kresin C, et al., 2024. Forecasting Doubling Time of SARS-CoV-2 using Hawkes and SQUIDER Models, Medical Research Archives, [online] 12(3). https://doi.org/10.18103/mra.v12i3.5137

DOI
https://doi.org/10.18103/mra.v12i3.5137

ISSN: 2375-1924

RESEARCH ARTICLE

Forecasting Doubling Time of SARS-CoV-2 using Hawkes and SQUIDER Models

Andrew Kaplan, UCLA Statistics

Conor Kresin, University of Otago Mathematics and Statistics

***Frederic Schoenberg**, UCLA Statistics

***Corresponding author:** frederic@stat.UCLA.edu

ABSTRACT
The rate of spread of an emerging epidemic is frequently characterized via the doubling time, which is the time it takes for the number of cases to double. This paper explores different ways to estimate doubling time, and investigates the estimation of doubling time in relationship to parameters in the HawkesN model and the SQUIDER (Susceptible, Quarantine, Undetected Infected, Infected, Dead, Exposed, Recovered) model. We observe an approximately exponential relationship between the productivity parameter κ in the HawkesN model and doubling time. We also evaluate the performance of the models in forecasting doubling times and compare to empirical doubling times using daily reported statewide totals for SARS-CoV-2 infections in California, and find that the HawkesN model forecasts doubling times more accurately, with 3.6% smaller root mean squared errors in Spring 2020, 79.4% smaller root mean squared errors in Autumn 2020, and 5.4% smaller root mean squared errors in Summer 2021. The HawkesN and SQUIDER models appear to forecast daily rate doubling times accurately at most times, though the SQUIDER forecasts of daily rate doubling times are far more volatile and thus occasionally have much larger errors, particularly in Fall 2020.

Keywords: Contagious diseases, epidemics, Hawkes model, Point process, SARS-Cov-2, Self-exciting.

1. Introduction

During the outbreak of an infectious disease, epidemiologists often characterize the rate of spread of the disease by discussing its doubling time. The term doubling time dates back to ancient Babylonian and Assyrian times to describe the growth of financial assets bearing interest, and has also been used frequently in population dynamics. The term is easily understood by the general public and is a useful summary of how quickly the population of infected individuals is expected to grow.

In the context of infectious diseases, however, there is some ambiguity in the definition of doubling time. Epidemiologists typically define doubling time as the time needed before the total cumulative number of cases is expected to double[1,2]. One may alternatively be interested in the time needed for the rate of daily new cases to double. We explore both here, and refer to the former as the cumulative doubling time and the latter as the rate doubling time. When the growth rate is exponential, the two constructs are identical, but when the growth of the epidemic is not perfectly exponential, various ambiguities and problems in estimating doubling time can arise.

The main questions explored in this paper are how accurately the doubling time can be estimated using two commonly used models for the spread of Covid-19 (SARS-CoV-2), and how this doubling time is related to parameters in these models. Despite the prevalence of the concept of doubling time in media reports and popular communication surrounding epidemic diseases such as Covid-19, surprisingly few studies have focused on doubling time in the scientific literature. A notable exception is the exploration of the relationship between doubling time and the reproduction number parameter, R_0, for the Susceptible-Infected-Recovered (SIR) model classically used in epidemiology to describe the spread of contagious diseases[3]. The SIR model may be overly simplistic, however, and presently, more sophisticated models are used to model Covid-19, such as the SQUIDER (Susceptible, Quarantine, Undetected Infected, Infected, Dead, Exposed, Recovered) model[4] and HawkesN model[5]. Here, we attempt to extend previous results on the SIR model and doubling time[3] by examining the relationship between doubling time and parameters in these more refined models for Covid-19, and we also consider the accuracy of estimates of doubling time using these models, employing data on confirmed SARS-CoV-2 cases in California during surges in 2020 and 2021. Studying doubling times using such a dataset is important not only for the purpose of understanding the relationship between doubling times and parameters in such models as HawkesN and SQUIDER which were used in forecasting the spread of Covid-19, but also to further our understanding of how to estimate doubling time and what level of accuracy can be anticipated for future epidemics.

The structure of the remainder of this paper is as follows. Following a brief description of HawkesN and SQUIDER models in Section 2, the California Covid-19 data used in the analysis are described in Section 3. Section 4 summarizes the results for estimating doubling time and its relation to HawkesN and SQUIDER parameters, and a discussion is given in Section 5.

2. HawkesN and SQUIDER Models.

2.1. THE HAWKESN MODEL

The self-exciting Hawkes model[6] is a commonly used point process model to describe clustering of random events occurring in time or space-time. Such point process models are typically characterized via their conditional intensity λ(t), which is the expected rate of occurrence of points around time t given information on all previously occurring points[7]. For the temporal Hawkes model, the conditional intensity is posited to obey

$$\lambda(t) = \mu + \kappa \sum_{i:t_i<t} g(t - t_i), \qquad (1)$$

where μ is the background rate at which points are immigrating into the current location, g is the triggering density governing the time it takes for one individual to infect another, and the parameter κ is called the productivity and t_i represents the expected number of people directly infected by each infected individual. Common choices for g are the exponential, Pareto, or normal densities, though sometimes non-parametric estimates are used. The parameter κ is closely connected to the reproduction number in compartmental models such as SIR. If $\kappa < 1$ is constant, then each point is expected to spread to $\kappa+\kappa^2 +\kappa^3 +\ldots = 1/(1-\kappa)-1 = \kappa/(1-\kappa)$ triggered points. As a result, in a Hawkes process with $\kappa < 1$, the expected fraction of background points is $1 - \kappa$. In recent applications to epidemic diseases[8,9], the productivity κ is typically allowed to vary over time, and thus is represented as a function κ(t).

Hawkes models have been used in a wide variety of applications including the forecasting of earthquakes[10,11], violent crimes[12,13] and the spread of epidemic diseases[14,15]. Such models have also been shown to be the best fitting models for forecasting seismicity in rigorous, purely prospective earthquake forecasting studies such as the Collaboratory for the Study of Earthquake Predictability (CSEP)[16–21].

Recent evidence has shown that Hawkes models, when fit to case counts of SARS-CoV-2 in the United States or Europe, Ebola in West Africa, or other epidemics, typically result in smaller forecast errors compared to alternative models such as compartmental models[22–25]. When used to model SARS-CoV-2 in the United States, Hawkes models resulted in a 31% reduction in root-mean-square (RMS) error, compared to SEIR models[24]. Further, Hawkes models and their variants such as the HawkesN model[5,26], and the recursive model[27] have been shown to be accurate for modeling not only SARS-CoV-2[8], but also Ebola [23,28] Chlamydia[29] SARS [30,31] measles[32], meningococcal disease[33], and Rocky Mountain Spotted Fever[27].

The idea behind the HawkesN model is that, as the number of previously infected individuals gets large relative to the total population size, the rate of spread of the disease should decrease due to herd immunity[5]. Thus, the HawkesN model has conditional intensity obeying

$$\lambda(t) = \left(1 - \frac{N_t}{N}\right)\left[\mu + \kappa \textstyle\sum_{i:\, t_i < t} g(t - t_i)\right], \qquad (2)$$

where the triggering density g is often chosen to be the exponential density function

$$g(u) = \beta e^{-\beta\mu} \qquad (3)$$

Here, N is the size of the population, and N_t is the number of individuals who have been infected prior to time t. The HawkesN model allows for the process to be non-stationary yet stable and non-explosive when κ > 1, since the rate of spread slows as the number of remaining susceptible individuals drops over time. The HawkesN model has been shown to be effective for forecasting statewide SARS-CoV-2 data in the United States[9], SARS-CoV-2 transmission in Indiana[8] and national SARS-CoV-2 data in the United States[25].

2.2. THE SQUIDER MODEL

Many compartmental models have been developed for describing the spread of epidemic diseases such as SARS-CoV-2. In the most basic compartmental model, the Susceptible Infected Recovered (SIR) model, the population is modeled as belonging to one of the three categories, and in each time unit, some proportion of the Susceptible population becomes Infected, and some proportion of the Infected proportion becomes Recovered. These proportions are typically modeled as fixed, i.e. not changing over time. A host of variants of the SIR model have been proposed, typically with more than just three compartments. One such variant is the Susceptible, Quarantine, Undetected Infected, Infected, Dead, Exposed, Recovered (SQUIDER) model, which has been shown to fit well and forecast accurately for the initial SARS-CoV-2 surge in the United States[4]. Like the SIR model, SQUIDER is a closed system of differential equations where individuals susceptible to the virus move from one compartment to another at fixed rates. The model is relatively simple, easily interpretable, and generally fits well to epidemic outbreaks, including SARS-CoV-2[24].

The SQUIDER model adds four additional compartments beyond what is present in the basic SIR model in order to take into account specifics regarding the behavior of the SARS-CoV-2 pandemic. The quarantine state (Q) takes into account subjects who have either been potentially exposed and quarantining at home for the required 10 days[34] or those who are staying at home voluntarily due to stay-at-home orders[4]. The undetected infected (U) and the undetected recovered / dead (E) states account for the fact that not all cases are detected, due to a lack of testing and unreported cases[4,35]. The dead (D) state represents the population of individuals who pass away from complications due to SARS-CoV-2[4]. The SQUIDER adaptation allows for a more accurate fit to SARS-CoV-2 cases in New York State than the basic SIR model, which is too simplistic to accurately predict various inflection points for confirmed infections[4].

The system of differential equations governing the SQUIDER model is thus as follows:

$$\frac{\partial S}{\partial t} = -\Theta S U^{a} - qS + \rho(E + R),$$
$$\frac{\partial U}{\partial t} = -\Theta S U^{a} - (q+\varepsilon+\delta)U,$$
$$\frac{\partial I}{\partial t} = \delta U - (\gamma + \alpha)I,$$
$$\frac{\partial R}{\partial t} = \alpha I - \rho R,$$
$$\frac{\partial D}{\partial t} = \gamma I,$$
$$\frac{\partial Q}{\partial t} = q(U + S),$$
$$\frac{\partial E}{\partial t} = \varepsilon U - \rho E,$$

where q = 0 except on days when quarantine periods (or stay-at-home orders) initiate or end.

3. Methods

3.1. PARAMETER ESTIMATION

When detailed occurrence times are available, the parameters in the Hawkes or HawkesN model are conventionally fit using maximum likelihood estimates, which are known to have desirable asymptotic properties including efficiency, consistency and asymptotic normality[36]. However, when only daily totals are available, as in the case for SARS-CoV-2 data in California, we use the least squares technique advocated in previous studies[9] based on the relationship observed between Hawkes processes and autoregressive time series[37,38].

Specifically, for the HawkesN model we find the least squares estimate

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\{\textstyle\sum_{t=1}^{T}(N(t) - [\mu + \kappa(t)\sum_{i:\, t_i < t} g(i)\, N(t - t_i)])^2\}, \qquad (4)$$

where $\theta = \{\mu,\kappa(t),\beta\}$, g obeys (3), T is the total length of the fitting period, $N(t)$ is the number of observed infections on day t, and $\kappa(t)$ is assumed to be constant within each surge but is permitted to vary from surge to surge. In other words, the parameters are chosen to minimize the sum of squared differences between the observed daily infection counts and the estimated case counts forecast using the HawkesN model (2). This least squares method is shown to be a reliable method of fitting Hawkes and HawkesN models in epidemiologic settings[8,9,25].

To fit the SQUIDER model, we follow the procedure recommended in previous studies, minimizing the sum of squared differences between expected and observed case and death counts[4]. The resulting fitted parameters include estimates of the rates of transmission (Θ), testing (δ), recovery (α), known deaths (γ), waning immunity (ρ) and undetected outcomes (ε) as well as the original proportion of undetected infected individuals, U(0), and U^a which allows the susceptible and undetected infectious populations to mix at a variable rate[4]. For the quarantine compartment in the SQUIDER model, both the time of the initial quarantine and the size of the population obeying the quarantine are to be estimated[4]. For this paper, since the quarantine origin and end times are known, we need only estimate the share of the population entering or leaving quarantine during each of the two stay-at-home orders issued in California[39,40] That is, we estimate q_1 = the number of Californians who began obeying the stay-at-home orders that commenced March 19th, 2020, q_2 = the number who ceased staying at home when the orders ended on May 18th, 2020, q_3 = the number obeying the stay-at-home orders beginning on November 19, 2020, and q_4 = the number who stopped following stay-at-home orders when they ended on January 25, 2021.

3.2. DATA

Records of California statewide totals of official daily reported cases of SARS-CoV-2 were obtained from the California Department of Public Health via their website, https://data.chhs.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state . The data were updated daily Mondays through Saturdays following their review and verification, and include all SARS-CoV-2 cases reported by state and territorial jurisdictions, including both confirmed and probable SARS-CoV-2 cases and deaths. Daily counts were obtained from February 16, 2020 through December 31, 2021, for a period of 684 days. Since doubling time is only defined during surges of an epidemic, we focus here on three surges: one beginning on 2/16/20, one beginning on 10/25/20, and one beginning on 6/20/21. To standardize the analysis we consider the first 31 days of each of these surges. The estimated total population of California was obtained from 2020 census records[40].

The dates corresponding to the recorded cases analyzed here may be quite different from the actual dates of onset of disease. Missing data are a serious potential problem with any study of SARS-CoV-2, as estimation of the number of unreported cases is exceedingly difficult [24,26]. A number of detailed studies were performed by the CDC in the Spring and Summer of 2020 in order to estimate the seroprevalence of the virus in several locations using sampling and testing of subjects at random[41]. Unfortunately, such careful studies ceased after the Trump administration cut funding for the CDC in summer 2020[42]. Further details on SARS-CoV-2 case surveillance data collection can be obtained from the California Department of Public Health[40] or CDC[2,43].

Of the three waves analyzed here, the first surge in Spring 2020 occurred when the California population 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[44]. During this time, testing for the virus was limited in California and there were major problems with testing[45]. The second surge during Autumn 2020 was also characterized by a dramatic increase in hospitalizations and fatalities statewide[2]. Although there had been some population exposure during Spring and Summer 2020, the proportion previously exposed was insufficient to provide general herd immunity[2,46]. The third wave during Summer 2021 occurred when the more infectious Delta variant of SARS-CoV-2 became dominant in California[47]. By this time, vaccines effective against the Delta strain such as BNT162b2 developed by Pfizer[48] had been approved for emergency use by the U.S. Food and Drug Administration for those over the age of 12 years[49]. However, only 68% of the eligible population in California was fully vaccinated by June 20, 2021[40] which did not provide for sufficient herd immunity against the Delta variant.

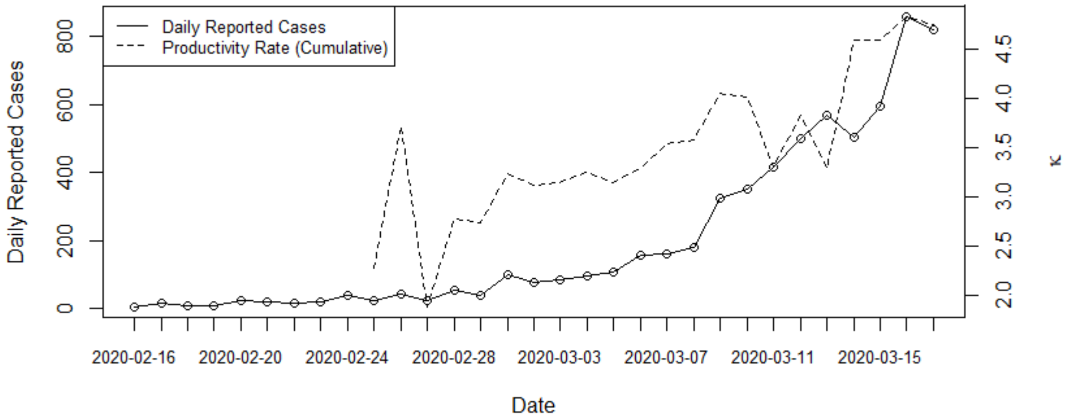
3.3. FORECASTING

The accuracy of doubling time estimates is considered for both retrospective and prospective analysis. That is, we assess the accuracy of doubling time estimates according to the fitted model with parameters estimated using data from the entire surge, and we also assess the accuracy of prospective, forecast doubling times, where for day t, the parameters in the model are fit using only observations up to and including day t. We then compare the observed doubling time with the median doubling time from simulations of the given HawkesN or SQUIDER model. In each case, the median of 100 simulated doubling times is used as the estimated doubling time, for both the rate doubling time and cumulative case doubling time, meaning the time elapsed until the daily number of cases at least doubles, and the time elapsed until the total cumulative number of cases at least doubles, respectively. HawkesN simulations are performed using the HawkesN Process Generator described in Appendix A, and SQUIDER simulations are performed using the method described in previous studies[4]. The root-mean-squared error (RMSE) is used to summarize the errors in the forecast doubling times based on the simulations of the models.

4. Results

Figure 1 shows how the estimated productivity parameter (κ) in the fitted HawkesN model evolves over time during each of the three surges. Along with the estimates of κ, the number of recorded cases per day of SARS-CoV-2 cases in California is also shown. In each of the three surges, the estimates of κ appear to stabilize after approximately 20 days. In the Spring 2020 surge, estimates of κ settled mostly between 3.3 and 4.6, whereas in the Fall 2020 and Summer 2021 surges, the estimates of κ were considerably lower, settling in the ranges of 1.8-2.2 and 2.0-2.7, respectively.

Figure 2 shows, based on simulations, how the cumulative and daily rate doubling times relate to the productivity parameter κ in the HawkesN model. The cumulative doubling time decreases approximately exponentially as κ increases, and this exponential decrease appears not to depend on the parameter β governing the transmission time density. The daily rate doubling time also appears to decrease roughly exponentially as κ increases, though the daily rate doubling times are considerably more noisy, as expected. For fixed κ, cumulative doubling times increase as β decreases, since β represents the inverse of generation length, so when β is very small, the disease is spreading more slowly. Both the cumulative doubling time and daily rate doubling time are calculated here as the time elapsed from the 50th recorded infection until the time of the corresponding doubling, as in the Johns Hopkins University and Medicine COVID-19 Dashboard[50].



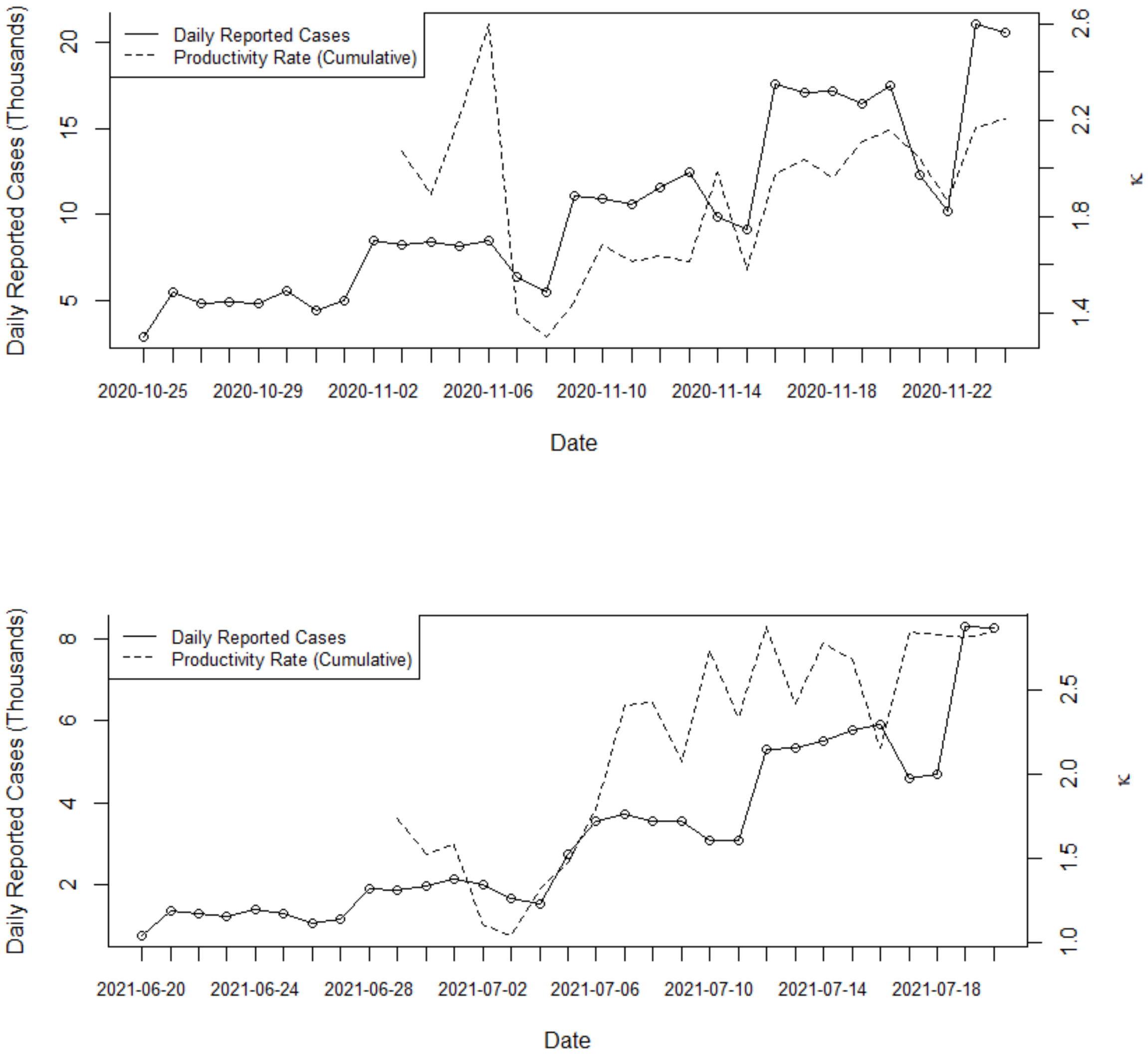
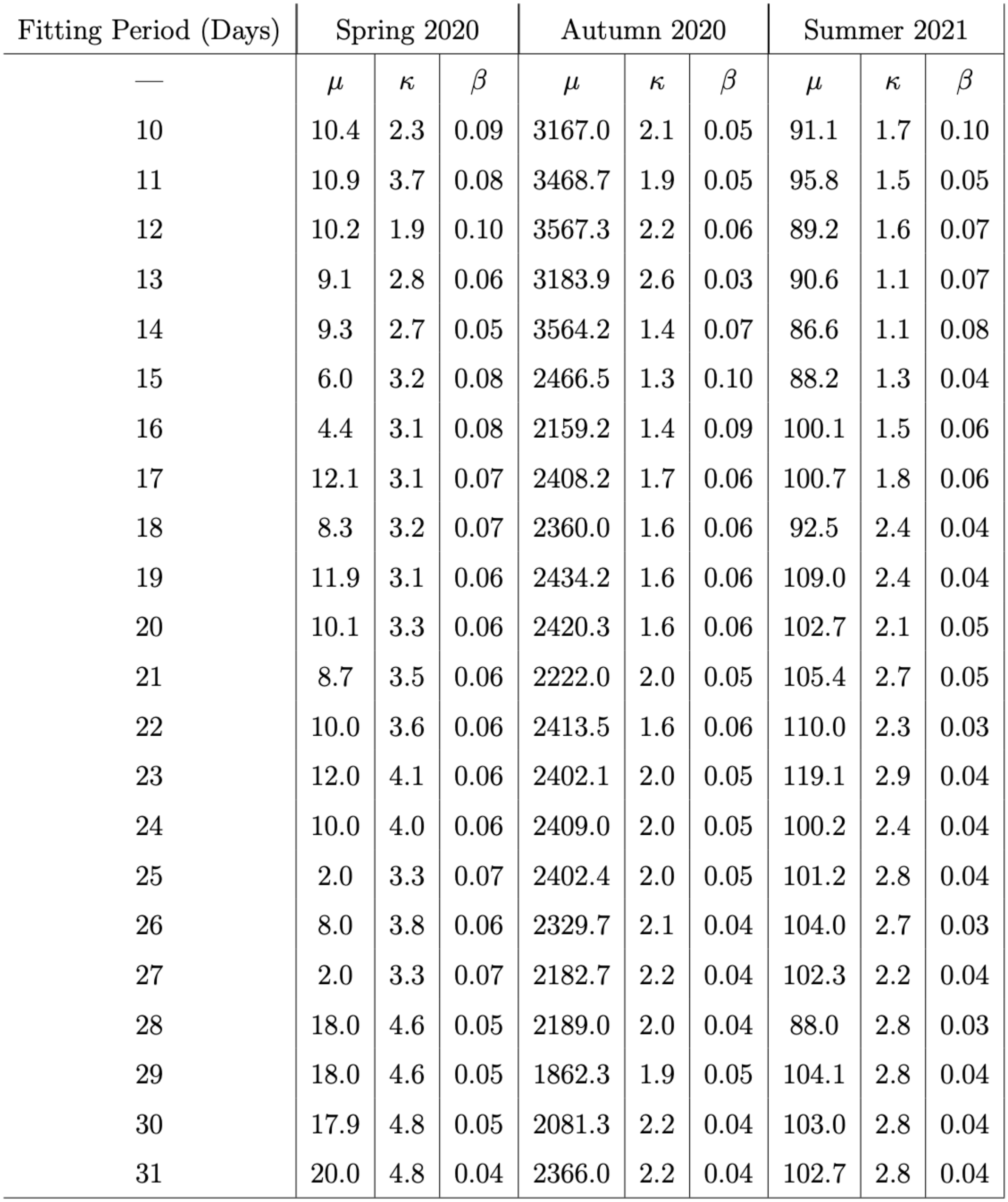


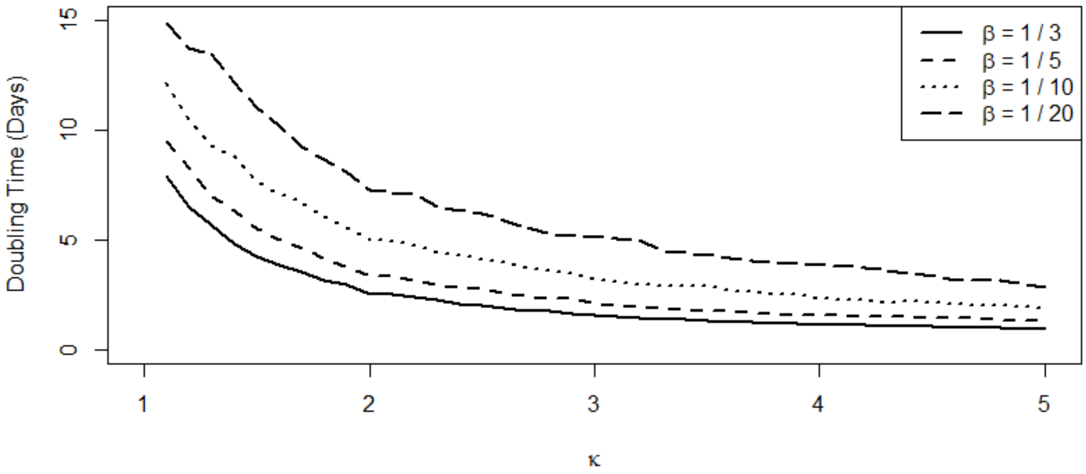
Figure 1: Daily number of reported cases and estimated productivity, $\hat{\kappa}$, for fitted HawkesN model, for Spring 2020 time period (top), Autumn 2020 (middle), Summer 2021 (bottom). For each day *t*, productivity is estimated using data from beginning of the plotted period up to and including day *t*.

Table 1 shows how the HawkesN parameter estimates, fit by least squares to the California SARS-CoV-2 data, vary as the length of the observation period increases. The estimated background rate μ was highest for the Autumn 2020 time period, whereas the estimates of κ were 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 version of the Hawkes model[27]. Estimates of β were generally higher during the initial Spring 2020 surge, perhaps due to a longer time period between exposure and symptomatic disease for the original variant[43,47].



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 1: Estimated HawkesN parameters μ, κ, β for training periods of varying lengths. Parameters are estimated by least squares.



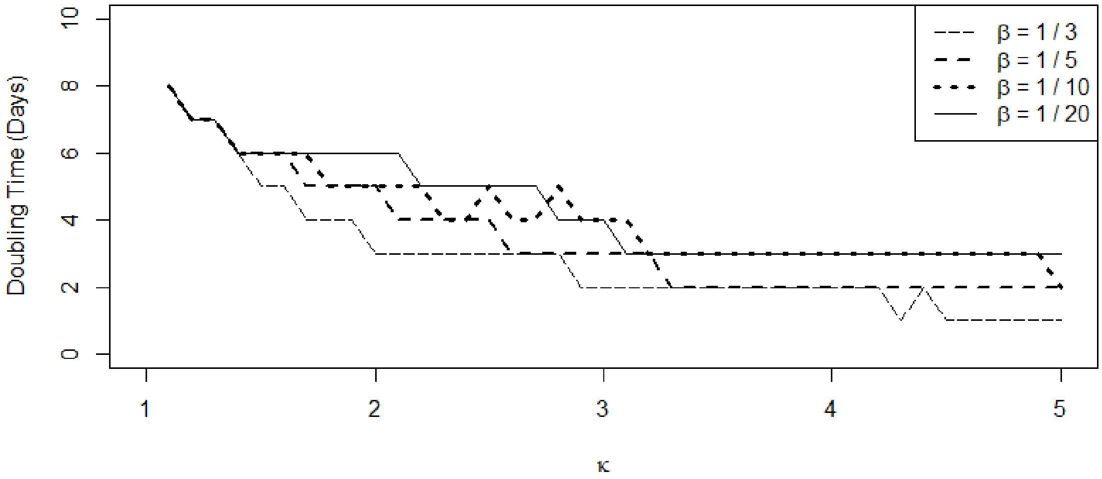
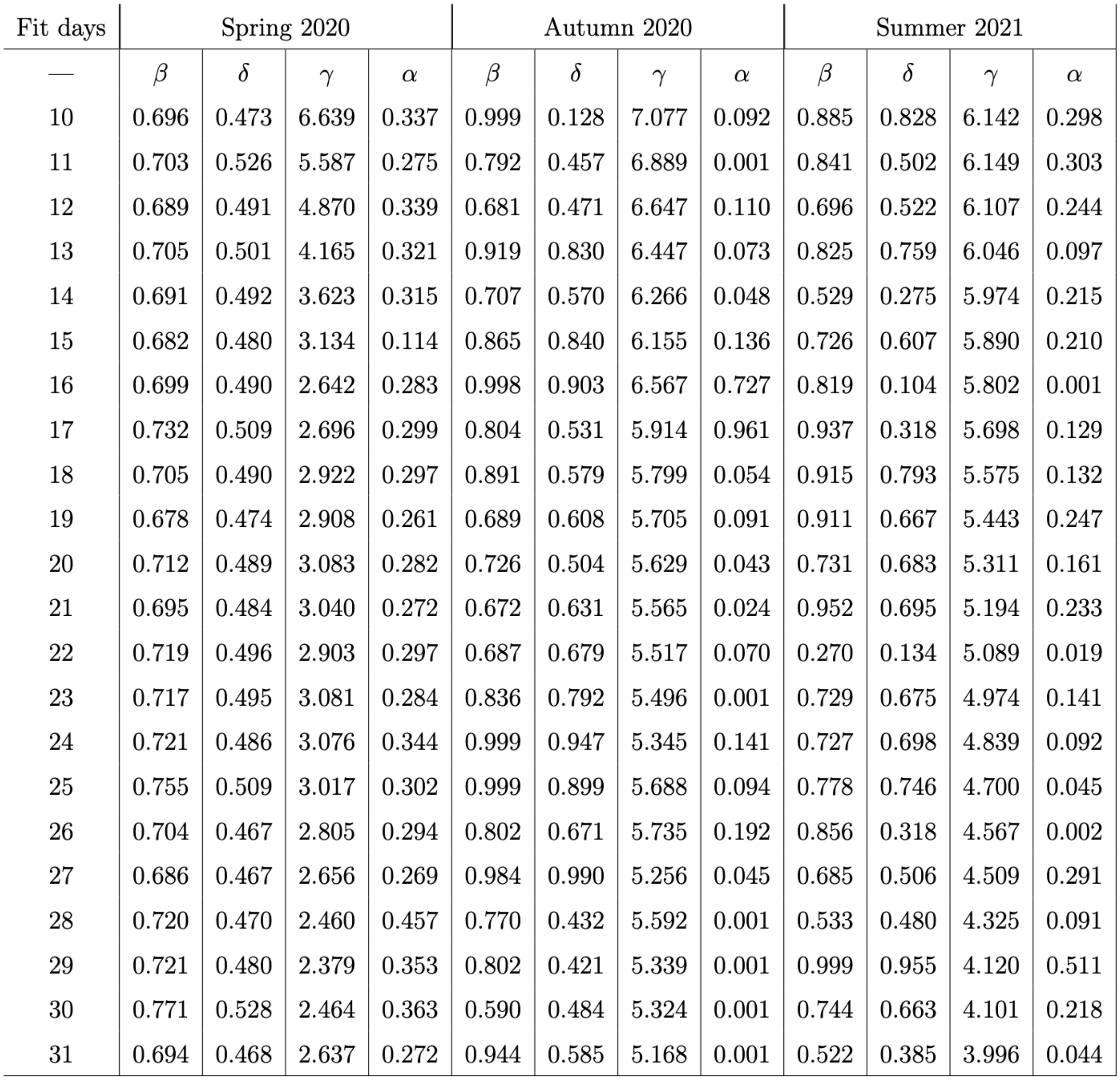


Figure 2: Doubling times over 500 simulations where β = {1/3, 1/5, 1/10, 1/20} . Top: median cumulative doubling time from the time when 50 total cases have been reported during the current surge. Bottom: median daily rate doubling time from the time when 50 total infections have been reported during the current surge.



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

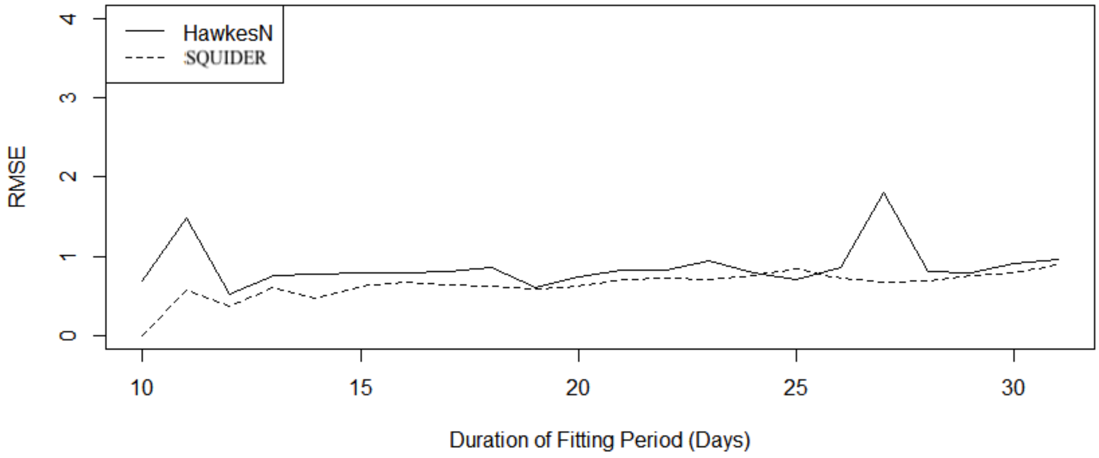
Table 2: Fitted values for selected parameters α, Θ, δ and γ in the SQUIDER model, estimated for training periods of varying lengths.

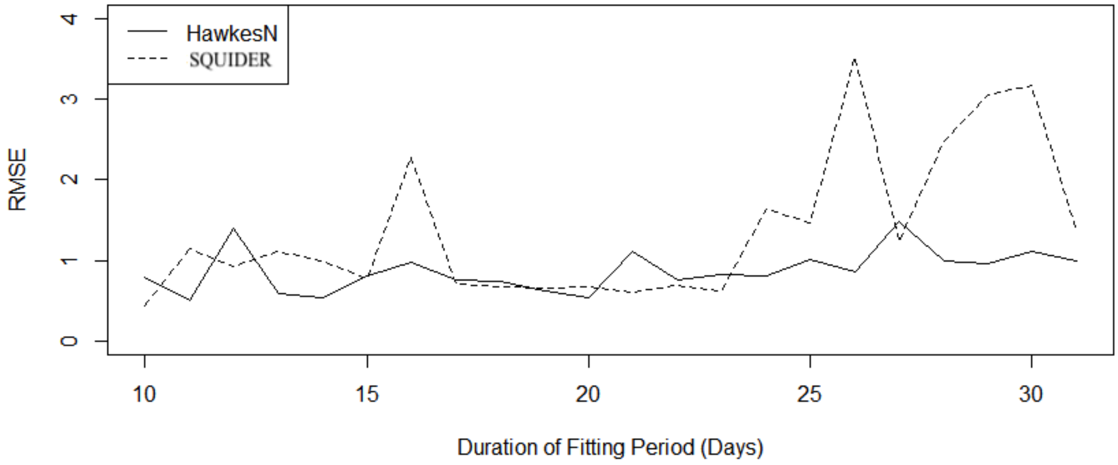
Table 2 provides estimates for selected parameters (Θ, δ, γ, α) for the SQUIDER model fit to the same California SARS-CoV-2 data. The estimated contact rate Θ is relatively constant for most fit lengths in all three time periods, ranging between 0.6 and 0.9. The estimated testing rate δ is generally higher during the Autumn 2020 and Summer 2021 SARS-CoV-2 surges than during Spring 2020. 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 Figures 3 and 4, the cumulative and rate doubling times for the fitted HawkesN and SQUIDER models are compared retrospectively, where the models are fit using the entire surge, and for each value of t, the simulated, model-based estimate of the doubling time is compared with the observed time required to at least double the number of cases having occurred on day t. For estimating cumulative doubling times, the HawkesN model has a higher retrospective RMSE than the SQUIDER model during Spring 2020 and a somewhat lower retrospective RMSE than the SQUIDER model during Fall 2020, especially for t in the range of 25 to 31 days. The retrospective RMSE of estimates of daily rate doubling times is higher for the HawkesN model than for the SQUIDER model especially during the Spring 2020 and Summer 2021 surges, particularly for t less than 26 days, though for Fall 2020 the HawkesN model's retrospective daily rate doubling times have lower RMSE than those of the SQUIDER model.

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: Root-mean-squared errors (RMSE), in days, for prospective, forecasted cumulative and daily rate doubling times for each model.





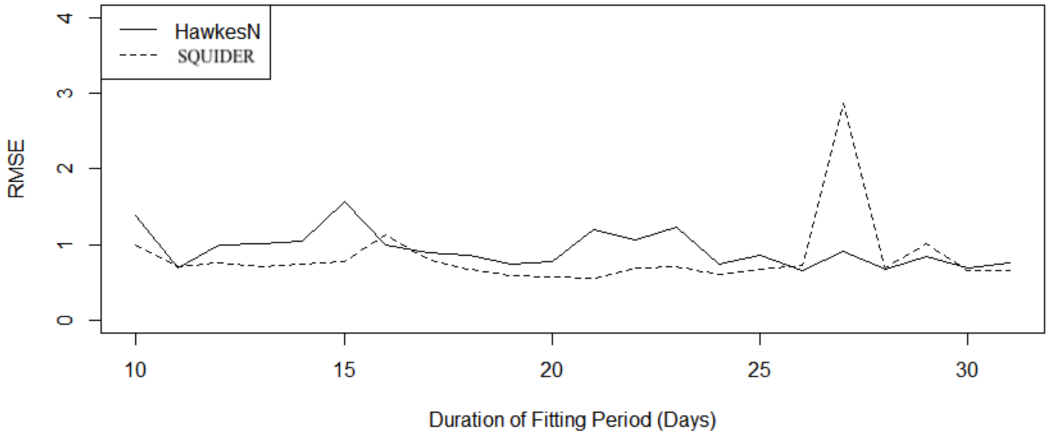
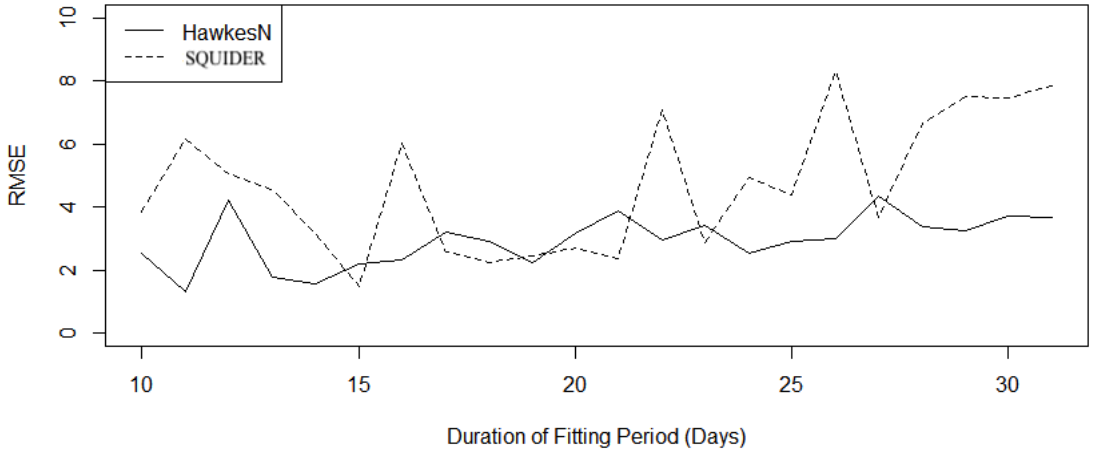
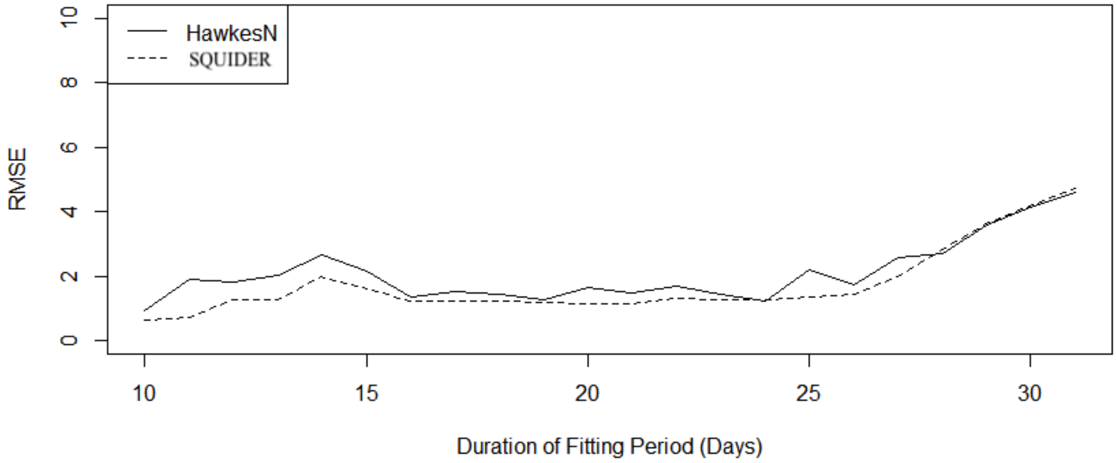


Figure 3: Root-mean-squared errors (RMSEs) of retrospective cumulative doubling time estimates for the HawkesN and SQUIDER models. Spring 2020 surge (top), Fall 2020 surge (middle), Summer 2021 surge (bottom).





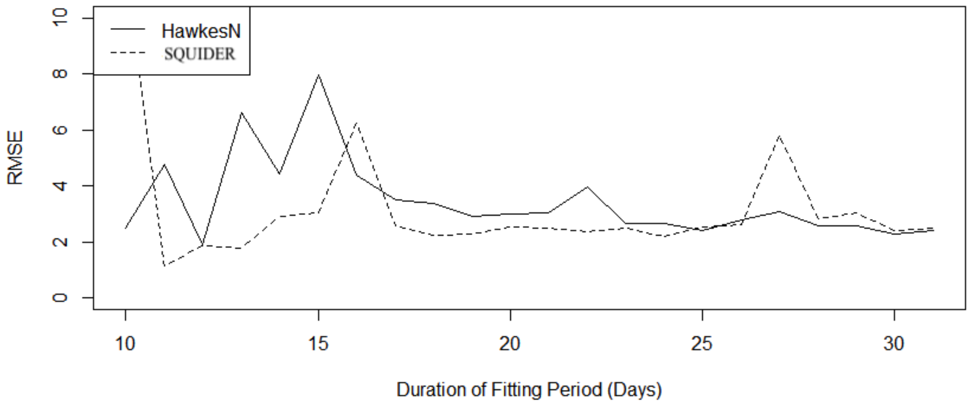
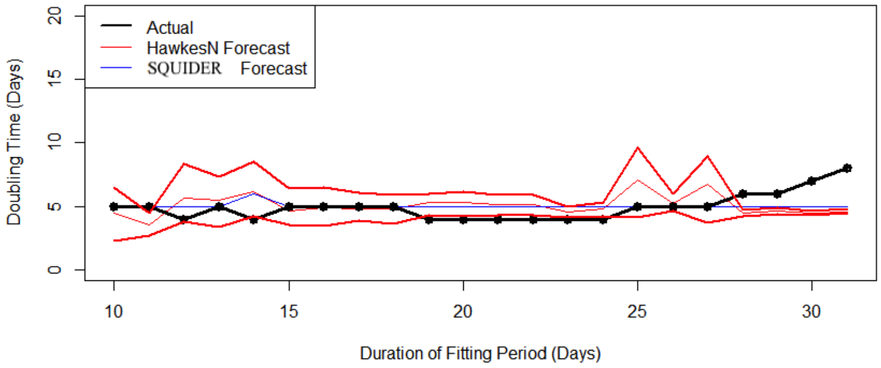
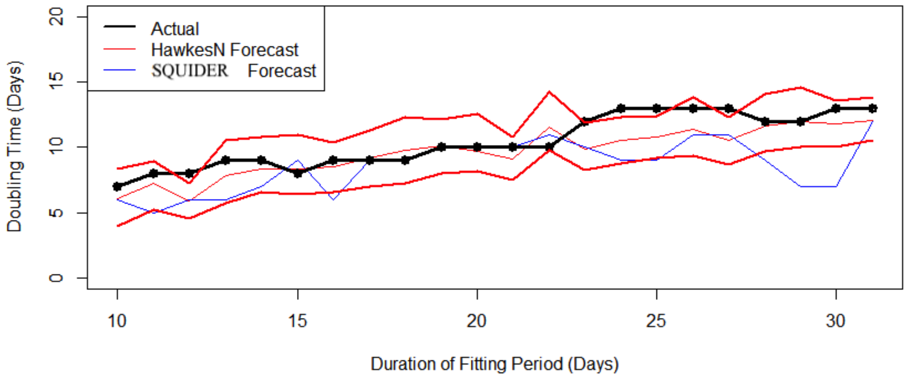


Figure 4: Root-mean-squared errors (RMSEs) of retrospective daily rate doubling time estimates for the HawkesN and SQUIDER models. Spring 2020 surge (top), Fall 2020 surge (middle), Summer 2021 surge (bottom).

However, as shown in Table 3, the prospective or forecasting doubling time estimates for the HawkesN model are more accurate overall than those of the SQUIDER model. An exception is Spring 2020, when the SQUIDER model estimates of cumulative doubling times have lower RMSE than those of HawkesN. However, in Fall 2020 and Summer 2021, the HawkesN estimates of cumulative doubling time are more accurate, and when estimating daily rate doubling times, the HawkesN estimates are more accurate than the SQUIDER estimates in all three surges. The daily rate doubling estimates for the HawkesN model are particularly more accurate than those for the SQUIDER model, with an RMSE of 4.251 days for the HawkesN forecast daily rate doubling times, compared to an RMSE of 20.621 days for the SQUIDER forecasts.





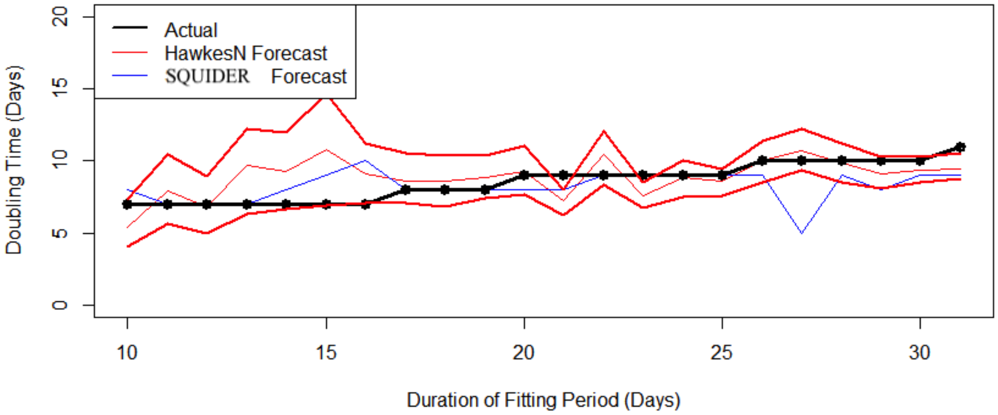


Figure 5: Prospective forecast cumulative doubling time estimates for the HawkesN model (red) and SQUIDER model (blue), along with the observed cumulative doubling time (black). For the HawkesN model, the thick red line represents the median of 100 simulations and the thin dashed red lines represent the middle 90% bounds based on the simulations. Spring 2020 surge (top), Fall 2020 surge (middle), Summer 2021 surge (bottom).

The forecast cumulative doubling times for the HawkesN and SQUIDER models are shown in Figure 5, along with the observed cumulative doubling times. 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 Fall 2020 SARS-CoV-2 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, with 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 6, both the HawkesN and SQUIDER models forecast daily rate doubling times accurately in most cases. However, the SQUIDER forecasts of daily rate doubling times appear to be far more volatile and thus occasionally have much larger errors, particularly in Fall 2020. However, both models underpredict the daily rate doubling time in late March and early April 2020, when the rate of daily new recorded infections slowed, as neither model was able to anticipate this change. During both Fall 2020 and Spring 2021, the middle 90% range of simulations of HawkesN forecasts of daily rate doubling times contain the observed daily rate doubling time in 39 out of these 44 forecasts.

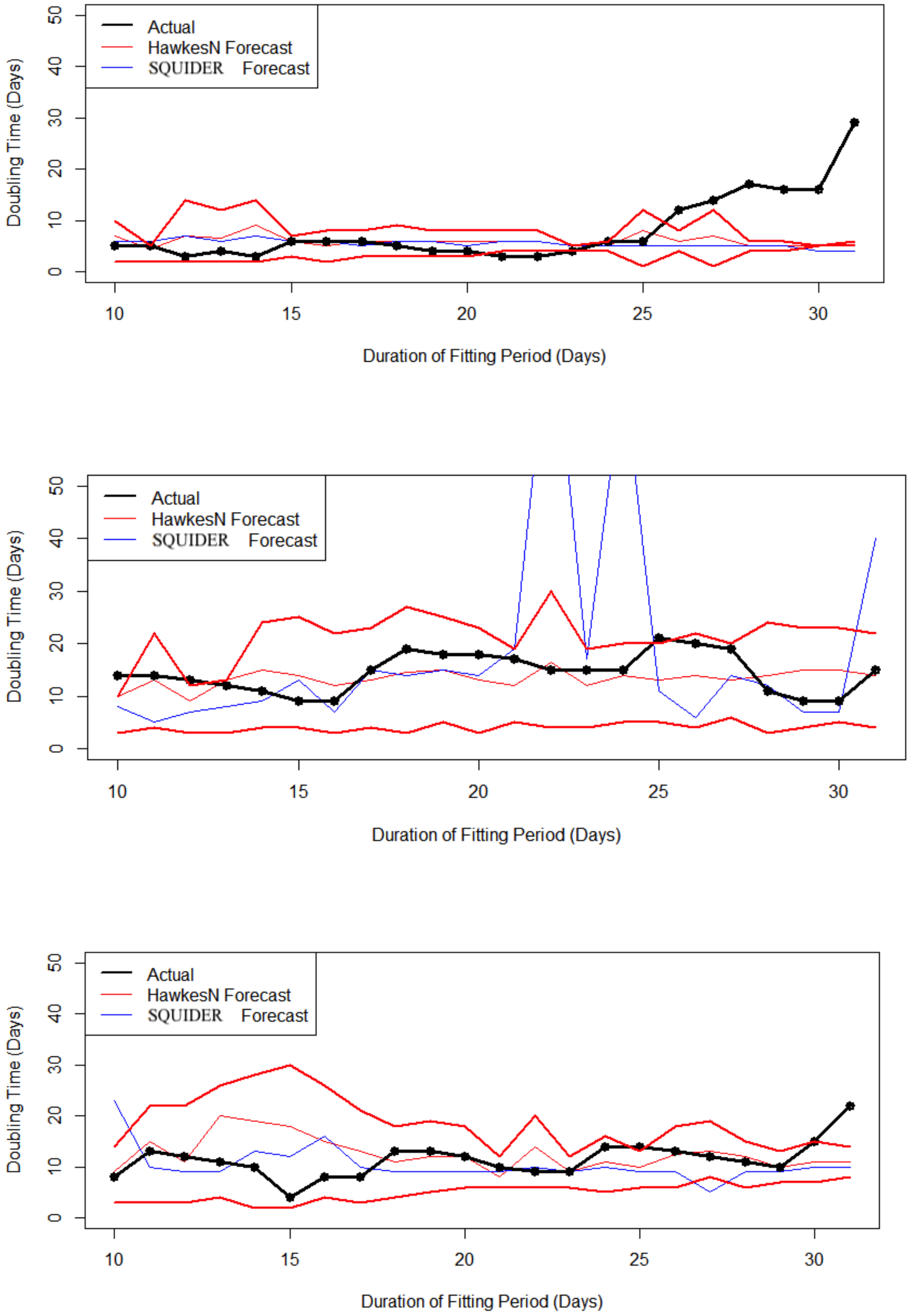


Figure 6: Prospective forecast daily rate doubling time estimates for the HawkesN model (red) and SQUIDER model (blue), along with the observed daily rate doubling time (black). For the HawkesN model, the thick red line represents the median of 100 simulations and the thin dashed red lines represent the middle 90% bounds based on the simulations. Spring 2020 surge (top), Fall 2020 surge (middle), Summer 2021 surge (bottom).

5. Discussion

The approximately exponential relationship between the productivity parameter κ in the HawkesN model and the corresponding doubling time is not surprising, as such an exponential relationship is consistent with the exponential growth characteristic of the HawkesN model as well as compartmental models. The finding here that the HawkesN model forecasts doubling times more accurately than the SQUIDER model, with 3.6% smaller root mean squared errors in Spring 2020, 79.4% smaller root mean squared errors in Autumn

2020, and 5.4% smaller root mean squared errors in Summer 2021, is somewhat surprising, however, given the prevalent use of SQUIDER and other compartmental models in modeling Covid-19. The HawkesN and SQUIDER models appear to forecast daily rate doubling times accurately at most times, though the SQUIDER forecasts of daily rate doubling times are far more volatile and thus occasionally had much larger errors, particularly in Fall 2020.

The higher RMSE in estimated cumulative doubling times and daily rate doubling times during the Spring 2020 surge for the HawkesN model compared to the SQUIDER model is likely attributable to the fact that the HawkesN model is simpler, with only three fitted parameters compared to twelve in the SQUIDER model. As a result of these extra fitted parameters, the SQUIDER model likely overfits, which would explain why the HawkesN model is substantially more accurate than the SQUIDER model for forecasting both daily and cumulative doubling times, particularly for the Fall 2020 wave of SARS-CoV-2.

An important item for future research would be to explore 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, other than those corresponding to previously reported cases. 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[5]. 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 add to model performance individually when applied to forecasting epidemic diseases including SARS-CoV-2. 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[47]. For this application, we 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, but a longer term forecast might benefit from taking 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[43] and should be added back in when immunity wanes[51] or a new variant reduces the effectiveness of such a vaccine[47]. Estimating these quantities might be difficult, however, and the results here suggest that, for estimating doubling times at least, the SQUIDER model may already be prone to overfitting and thus yield larger errors compared to the simpler HawkesN model with fewer estimated parameters.

Acknowledgements. This material 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. Computations were performed in R.

References.

1. Centers for Disease Control and Prevention. 2009 H1N1 Early Outbreak and Disease Characteristics. https://www.cdc.gov/h1n1flu/surveillanceqa.htm. Published 2009. Accessed Jan, 2022.
2. Centers for Disease Control and Prevention. Coronavirus Disease 2019 (COVID-19). https://www.cdc.gov/dotw/covid-19/index.html. Published 2021. Accessed Jan, 2022.
3. Merler S, Ajelli M, Fumanelli L, Vespignani A. Containing the accidental laboratory escape of potential pandemic influenza viruses. BMC Medicine 2013; 11:252.
4. Khan ZS, Van Bussel F, Hussain F. A predictive model for Covid-19 spread - with application to eight US states and how to end the pandemic. Epidemiology & Infection 2020;148:e249. doi: 10.1017/S0950268820002423.
5. Rizoiu MA, Mishra S, Kong Q, Carman M, Xie L. SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Processes. Proceedings of the 2018 World Wide Web Conference 2018; 419-428.
6. Hawkes A. Spectra of some self-exciting and mutually exciting point processes. Biometrika 1971; 58(1): 83-90.
7. Daley DJ, Vere-Jones D. An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods, 2nd ed. 2003; Springer, New York.
8. Mohler G, Schoenberg F, Short MB, Sledge D. 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 2021; 8(1): 1-8.
9. Schoenberg F. Estimating Covid-19 transmission time using Hawkes point processes. Annals of Applied Statistics, 2023; 17(4): 3349-3362.
10. Ogata Y. Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association 1988; 83(401): 9-27.
11. Ogata Y. Space-time point process models for earthquake occurrences. Ann. Inst. Statist. Math 1998; 50(2): 379-402.
12. Mohler GO, Short MB, Brantingham, PJ, Schoenberg FP, Tita GE. Self-exciting point process modeling of crime. J. Amer. Statist. Assoc. 2011; 106(493): 100-108.
13. Park J, Schoenberg FP, Brantingham PJ, Bertozzi AL. Investigating clustering and violence interruption in gang-related violent crime data using spatial-temporal point processes with covariates. J. Amer. Statist. Assoc. 2021; 116 (536): 1674-1687.
14. Meyer S, Elias J, Hohle M. A space-time conditional intensity model for invasive meningococcal disease occurrence. Biometrics 2012; 68(2): 607-616.
15. Meyer, S., Held L. (2014), Power-law models for infectious disease spread. AoAS, 2014; 8(3): 1612-1639.
16. Clements RA, Schoenberg FP, Schorlemmer D. Residual analysis for space-time point processes with applications to earthquake forecast models in California. Annals of Applied Statistics 2011; 5(4): 2549-2571.
17. Clements RA, Schoenberg FP, Veen A. Evaluation of space-time point process models using super-thinning. Environmetrics 2013; 23(7): 606-616.
18. Zechar JD, Schorlemmer D, Werner MJ, Gerstenberger MC, Rhoades DA, Jordan TH. Regional Earthquake Likelihood Models I: First-order results. Bull. Seismol. Soc. Am. 2013; 103: 787-798.
19. Bray A, Wong K, Barr CD, Schoenberg FP (2014). Voronoi cell based residual analysis of spatial point process models with applications to Southern California earthquake forecasts. Annals of Applied Statistics 8(4), 2247-2267.
20. Gordon JS, Clements RA, Schoenberg FP, Schorlemmer D. Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. Spatial Statistics 2015; 14:133-150.
21. Schorlemmer D, Werner MJ, Marzocchi W, Jordan TH, Ogata Y, Jackson DD, Mak S, Rhoades DA, Gerstenberger MC, Hirata N, Liukis M, Maechling PJ, Strader A, Taroni M, Wiemer S, Zechar JD, Zhuang J. The Collaboratory for the Study of Earthquake Predictability: achievements and priorities. Seismological Research Letters 2018; 89(4):1305-1313.
22. Yang AS. Modeling the Transmission Dynamics of Pertussis Using Recursive Point Process and SEIR model. PhD thesis, UCLA, 2019. Los Angeles, CA.
23. Park J, Chaffee A, Harrigan R, Schoenberg FP. A non-parametric Hawkes model of the spread of Ebola in West Africa. J Appl. Stat. 2022; 49(3): 621-637.
24. Kresin C, Schoenberg F, Mohler G. Comparison of Hawkes and SQUIDER models for the spread of Covid-19. Advances and Applications in Statistics 2021; 74: 83-106.
25. Chiang WH, Liu X, Mohler G. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. Int. J. Forecast. 2022; 38(2): 505-520.
26. Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proceedings

of the National Academy of Sciences 2020; 117(29): 16732-16738.

27. Schoenberg FP, Hoffmann M, Harrigan, R. A recursive point process model for infectious diseases. AISM 2019; 71(5): 1271-1287. https://doi.org/10.1007/s10463-018-0690-9.
28. Kelly JD, Harrigan RJ, Park J, Hoff NA, Lee SD, Wannier R, Selo B, Mossoko M, Njokolo B, Okitolonda-Wemakoy E, Mbala-Kingebeni P, Rutherford GW, Smith TB, Ahuka-Mundeke S, Muyembe-Tamfum JJ, Rimoin AW, Schoenberg FP.
    Real-time predictions of the 2018-2019 Ebola virus disease outbreak in the Democratic Republic of Congo using Hawkes point process models. Epidemics 2019; 28, 100354. doi: 10.1016/j.epidem.2019.100354.
29. Schoenberg F. Nonparametric estimation of variable productivity Hawkes processes. Environmetrics 2022; 33(6): e2747.
30. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Amer. J. Epidemiology 2004; 160(6):509-516.
31. Cauchemez S, Boelle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, Hedley AJ, Anderson RM, Valleron AJ. Real-time estimates in early detection of SARS. Emerging infectious diseases 2006; 12(1):110.
32. Farrington C, Kanaan M, Gay N. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 2003; 4(2):279-295.
33. Meyer S, Held L, Hohle, M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. Journal of Statistical Software 2017; 77 (11): 1-55.
34. Centers for Disease Control and Prevention. Quarantine and Isolation. https://www.cdc.gov/coronavirus/2019-ncov/your-health/quarantine-isolation.html. Published 2021. Accessed Dec, 2021.
35. Bŏhning D, Rocchetti I, Maruotti A, Hollinge H. Estimating the undetected infections in the Covid-19 outbreak by harnessing capture-recapture methods. Int. J. Infect. Dis. 2020; 97: 197-201.
36. Ogata Y. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. Annals of the Institute of Statistical Mathematics 1978; 30(2): 243-261.
37. Kirchner M. Hawkes and INAR(∞) processes. Stochastic Processes and their Applications 2016; 26(8): 2494-2525.
38. Kirchner M. An estimation procedure for the Hawkes process. Quant. Financ. 2017; 17(4):571-595.
39. Newsom G, Padilla A. Executive Order N-33-20. Executive Department State of California. https://covid19.ca.gov/img/Executive-Order-N-33-20.pdf. Published 2020. Accessed Jan, 2022.
40. California Department of Public Health. 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 . Published 2022. Accessed Aug, 2021.
41. Bajema KL, Wiegand RE, Cuffe K, Patel SV, Iachan R, Lim T, Lee A, Moyse D, Havers F, Harding L, Fry AM, Hall AJ, Martin K, Biel M, Deng Y, Meyer WA, Mathur M, Kyle T. Estimated SARS-CoV-2 Seroprevalence in the US as of September 2020. JAMA Intern Med. 2021; 181(4):450-460.
42. Wermer E, Stein J. Trump administration pushing to block new money for testing, tracing and CDC in upcoming coronavirus relief bill. Washington Post, 2020;July 18: 1-2.
43. Centers for Disease Control and Prevention. COVIDView – Key Updates for Week 3, ending January 30, 2021. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html. Published 2021. Accessed Jan, 2022.
44. Hogan C, Garamani N, Sahoo M, Huang CH, Zehnder J, Pinsky B. Retrospective Screening for SARS-COV-2 RNA in California, USA, Late 2019. Emerg Infect Dis 2020; 26(10): 2487-2488.
45. Patel, N. (2020). Why the CDC botched its coronavirus testing. MIT Technology Review. www.technologyreview.com/2020/03/05/905484/ why-the-cdc-botched-its-coronavirus-testing . Published Mar 5, 2020. Accessed Jan, 2022.
46. Fontanet A, Cauchemez S. COVID-19 herd immunity: where are we? Nat Rev Immunol. 2020; 20(10): 583-584.
47. Centers for Disease Control and Prevention. Delta Variant: What We Know About the Science. https://www.cdc.gov/coronavirus/2019-ncov/variants/about-variants.html. Published 2021. Accessed Dec, 2021.
48. BianL,GaoQ,GaoF,WangQ,HeQ,WuX,MaoQ,XuM,LiangZ. Impact of the Delta variant on vaccine efficacy and response strategies. Expert Rev. Vaccines 2021; 20(10): 1201-1209.
49. U.S. Food & Drug Administration (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. https://www.fda.gov/news-events/press-announcements/ coronavirus-covid-19-update-fda-authorizes-pfizer-biontech-covid-19vaccine-emergency-use. Published May 10, 2021. Accessed Jan, 2022.
50. Johns Hopkins University and Medicine. Cumulative Cases by Days since 50th Confirmed

Case, Coronavirus Resource Center. https://coronavirus.jhu.edu/data/cumulative-cases . Published 2022. Accessed Jan, 2022.

51. Hamady A, Lee J, Loboda Z. Waning antibody responses in COVID-19: what can we learn from the analysis of other coronaviruses? Infection 2022;50(1): 11-25.

52. Lewis PAW, Shedler GS. Simulation of nonhomogeneous poisson processes by thinning. Naval Research Logistics Quarterly 1979; 26(3), 403-413.

6. Appendix

6.1. 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. 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[24]. The conditional intensity of new infections is given by

$$\lambda^{E}(t) = (1 - \frac{N^{E}{}_{t}}{N}) \sum_{t>t_j{}^{I}} R_0\, \gamma\, exp\{-\gamma(t - t_j{}^{I})\}, \qquad (5)$$

and infecton times are generated by

$$P(t_j{}^{I} > t_j{}^{E} + c) = \int_c^{\infty} \mu\, exp(-\mu(s - t_j{}^{E}))ds. \qquad (6)$$

In (5), the conditional intensity is still a function of the susceptible population, productivity and triggering function as in (2), 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 (6), a new infection t^{I}_{j} 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 (2)). The SEIR-Hawkes process can be simulated using an iterative process as shown in previous studies[24].

In this paper, we develop an algorithm to simulate HawkesN processes using a similar method to the one used in prior studies[24]. 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 = κ and γ = β. 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:

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, offset the time of exposure from the ancestor's by *exp(μ).* Last, sort in chronological order all of the accepted points and all of the candidate points together. 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*:

$$\lambda_c(t) = (1 - \frac{N_a}{N}) \sum_{i=1}^{N_a} R_0\, \gamma\, exp\{-\gamma[t(c) - t(aj)]\}, \qquad (7)$$

and

$$\nu_c = \sum_{i=1}^{N_S} R_0\, \gamma\, exp\{-\gamma[t(c) - t(sj)]\}, \qquad (8)$$

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[52]. That is, accept the candidate point if

$$D_c \sim \text{Unif}[0, 1] < \frac{\lambda_c}{\nu_c}. \qquad (9)$$

Part 4: Finally, to take into account exposure time before infection, the time of infection for each newly accepted point is offset by an exponential amount:

$$t(a) = t(a) + exp(\gamma). \qquad (10)$$