Submitted to the Annals of Applied Statistics

ESTIMATING COVID-19 TRANSMISSION TIME USING HAWKES POINT PROCESSES

BY FREDERIC SCHOENBERG[1,a],

[1]*Department of Statistics, University of California, Los Angeles* , [a]*frederic@stat.UCLA.edu*

The question addressed here is whether the distribution of SARS-CoV-2 (Covid-19) transmission times can be estimated accurately with only case count data, using Hawkes models. We fit Hawkes models with varying productivities to each of the 50 United States individually, estimating for each state a transmission time density, both non-parametrically and using a normal approximation. We find that for nearly all states, the estimated transmission times are centered near 7 days with a standard deviation of approximately 1 day. Compared to previous reports, the results here suggest that transmission times for SARS-CoV-2 are somewhat shorter on average and the distribution is less diffuse, though the results also suggest the possibility of transmission occurring on the first day of exposure.

1. Introduction. An important problem in the statistical modeling of the spread of SARS-CoV-2 (also commonly known as Covid-19) in the United States is the estimation of the distribution of transmission times, and there remains considerable uncertainty about this distribution. The transmission time, or time elapsed between when a person is infected and when that person infects someone else, is closely related to the incubation period, which is the time between exposure to the disease until the time the disease becomes symptomatic. A variety of medical reports and early case studies have investigated the typical incubation duration for SARS-CoV-2, finding that the incubation period for SARS-CoV-2 has a median time of 4-5 days but can be up to 14 days (Guan et al. 2020, Lauer et al. 2020, Li et al. 2020), with 97.5% of symptomatic patients with SARS-CoV-2 exhibiting symptoms within 11.5 days of SARS-CoV-2 infection (Lauer et al. 2020). In addition, early studies of patients in Wuhan, China reported that the median time from the onset of SARS-CoV-2 illness to acute respiratory distress syndrome (ARDS) was 8-12 days and the median time from onset of illness to ICU admission was 9.5-12 days (Huang et al. 2020, Wang et al. 2020, Yang et al. 2020, Zhou et al. 2020). Based on these studies, the Centers for Disease Control and Prevention (CDC) summarized that the incubation period ranges from 2-14 days (CDC 2021a), that the contagious period for those with SARS-CoV-2 is typically up to 10 days following symptom onset (CDC 2021d) or 14 days following exposure (CDC 2021c,e), but may extend up to 20 days after exposure (CDC 2021a), and recommended infected individuals stay home for 14 days after last contact with someone with SARS-CoV-2 (CDC 2021b,c). Similarly, the World Health Organization reported that because the incubation period averages 5-6 days but can range up to 14 days, infected individuals should quarantine for 14 days after exposure (WHO 2021a,b).

While these recommendations may be sensible and practical from public health and policy perspectives, from a statistical modeling point of view, a better understanding of the empirical distribution of transmission times is desirable. Such statistical models may be useful for forecasting, planning and mitigation of the pandemic, and typically require an accurate estimation of the distribution of the time between exposure and transmission of the virus. Since

Keywords and phrases: Disease epidemics, Hawkes model, Point process, Self-exciting.

transmission time ultimately depends critically on human behavior, which can vary significantly by location, it is important to complement biological experiments and case studies from Wuhan with studies of the empirical time between transmissions.

One way to investigate the transmission time for SARS-CoV-2 is by using daily reported case counts; with sufficient data and a sufficiently accurate statistical model for the spread of the disease, estimation of the transmission time may be feasible. Several different frameworks have been proposed for modeling the spread of SARS-CoV-2, including compartmental models such as the SEIR (Susceptible $\rightarrow$ Exposed $\rightarrow$ Infectious $\rightarrow$ Removed) differential equation model, and branching point process models such as the Hawkes point process model (Rizoiu et al. 2018, Jewell et al. 2020, Bertozzi et al. 2020, Chiang et al. 2020). Relative to Hawkes models, SEIR models and their variants have been used more widely to describe the SARS-CoV-2 pandemic (Bertsimas 2020, IHME 2020, You et al. 2020, LANL 2020) as well as other infectious diseases such as Ebola (Lekone and Finkenstadt 2006) and SARS-Cov-1 (Dye and Gay 2003). However, such compartmental models can have serious limitations when used to describe the detailed local behavior of an epidemic (Meyers 2007) and can significantly overpredict counts of infections such as SARS-Cov-1 (Xu et al. 2004). Recent evidence has shown that Hawkes models, when fit to case counts of SARS-CoV-2 or other epidemics, typically result in smaller forecast errors compared to alternative models such as SEIR or SVEILR models (Yang 2019, Kresin et al. 2021). 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 (Kresin et al. 2021). Further, Hawkes models and their variants such as the HawkesN model (Rizoiu et al. 2018, Bertozzi et al. 2020), and the recursive model (Schoenberg et al. 2019) have been shown to be accurate for modeling not only SARS-CoV-2 (Mohler et al. 2021), but also Ebola (Kelly et al. 2019, Park et al. 2020), Chlamydia (Schoenberg 2022), SARS (Wallinga and Teunis 2004, Cauchemez et al. 2006), measles (Farrington et al. 2003), meningococcal disease (Meyer et al. 2017), and Rocky Mountain Spotted Fever (Schoenberg et al. 2019). 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) (Clements et al. 2011, Clements et al. 2013, Zechar et al. 2013, Bray et al. 2014, Gordon et al. 2015, Schorlemmer et al. 2018).

Here, the emphasis is on estimating the transmission time distribution, and we use Hawkes models, estimated by least squares, fit to data on confirmed SARS-CoV-2 cases from each of the 50 United States. The structure of the remainder of this paper is as follows. Following a brief description of the data in Section 2, Hawkes processes and methods for their estimation and analysis are presented in Section 3, and Section 4 describes simulations confirming the accuracy of the estimation procedure. Section 5 summarizes the results, and a discussion is given in Section 6.

2. Data. Records of daily totals of CDC SARS-CoV-2 case surveillance data in each of the 50 United States were obtained from the CDC via their website, https://covid.cdc.gov/covid-data-tracker (CDC 2021f). For each state, we obtained daily counts from 1/23/20 to 08/25/21, for a period of 582 days. In what follows, because we estimated parameters over 36 windows each of length 16 days, we used $36 \times 16 = 576$ days in the analysis, corresponding to the dates 1/23/20 to 08/19/21. The data were publicly available and downloaded on 8/26/21.

Figure 1a shows the total number of recorded cases for each state during this 582 day reporting period. The totals closely follow the United States population, with California,

Texas and Florida having the most recorded cases, in line with their population sizes. Figure 1b displays the total cases per capita over this time period, using population data obtained at https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html . These publicly available data were downloaded on 1/21/23. The states with the highest incidence per capita were North Dakota (0.1489), Rhode Island (0.1471), South Dakota (0.1468), and Arkansas (0.1465). Most states had quite similar per capita rates of recorded incidence; 41 of the 50 states had per capita incidence between 0.10 and 0.15. Four states had very low per capita incidence over this recording period: these were Hawaii (0.0380), Vermont (0.0401), Maine (0.0548), and Oregon (0.0625).

As indicated on the CDC website, the data were updated daily Mondays through Saturdays following their review and verification, and may deviate from data on jurisdictional websites according to the timing of reporting of cases to the CDC. The totals include all SARS-CoV-2 cases reported by state and territorial jurisdictions to the CDC, with the exception of persons repatriated to the United States from Wuhan, China, and Japan (CDC 2021f), and for many jurisdictions the totals include both confirmed and probable SARS-CoV-2 cases and deaths. The recorded number of cases corresponding to a given day reflects information provided by the states and jurisdictions, and may thus represent either the date the case occurred or the date it was recorded in the state (CDC, 2021).

The dates corresponding to the recorded cases analyzed here may be quite different from the actual dates of onset of disease. What is considered a *transmission time* in the results and analysis below is really the time elapsed between recorded dates of two cases, and this time interval may reflect not only the time of incubation and expression of disease but also the difference in times for the cases to be recorded. If, for example, transmissions shortly after exposure are more likely to result in recorded cases than transmissions many days after exposure, then the mean transmission time may be underestimated.

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 (Bertozzi et al. 2020, Kresin 2021). 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 (Bajema et al. 2021). Unfortunately, such careful studies ceased after the Trump administration cut funding for the CDC in summer 2020 (Wermer and Stein 2020). Different states may have somewhat different rates of missing data (Bajema et al. 2021), though we are unaware of any particularly notable examples of states with especially unreliable data. For more information on the collection of the CDC SARS-CoV-2 case surveillance data, see https://www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance.html or CDC (2021).

Figure 2 shows the trend in reported cases per capita in six typical states. One sees the sharp, approximately exponential increases and decreases in incidence characteristic of epidemic models such as SEIR and Hawkes models (Rizoiu et al. 2018, Kresin et al. 2021).

3. Methods.

3.1. *Hawkes point process models.*. The Hawkes or *self-exciting* point process model (Hawkes 1971) is commonly used to model clustered point patterns in applications such as seismology, finance, crime, and infectious diseases (Ogata 1988, Daley and Vere-Jones 2003,

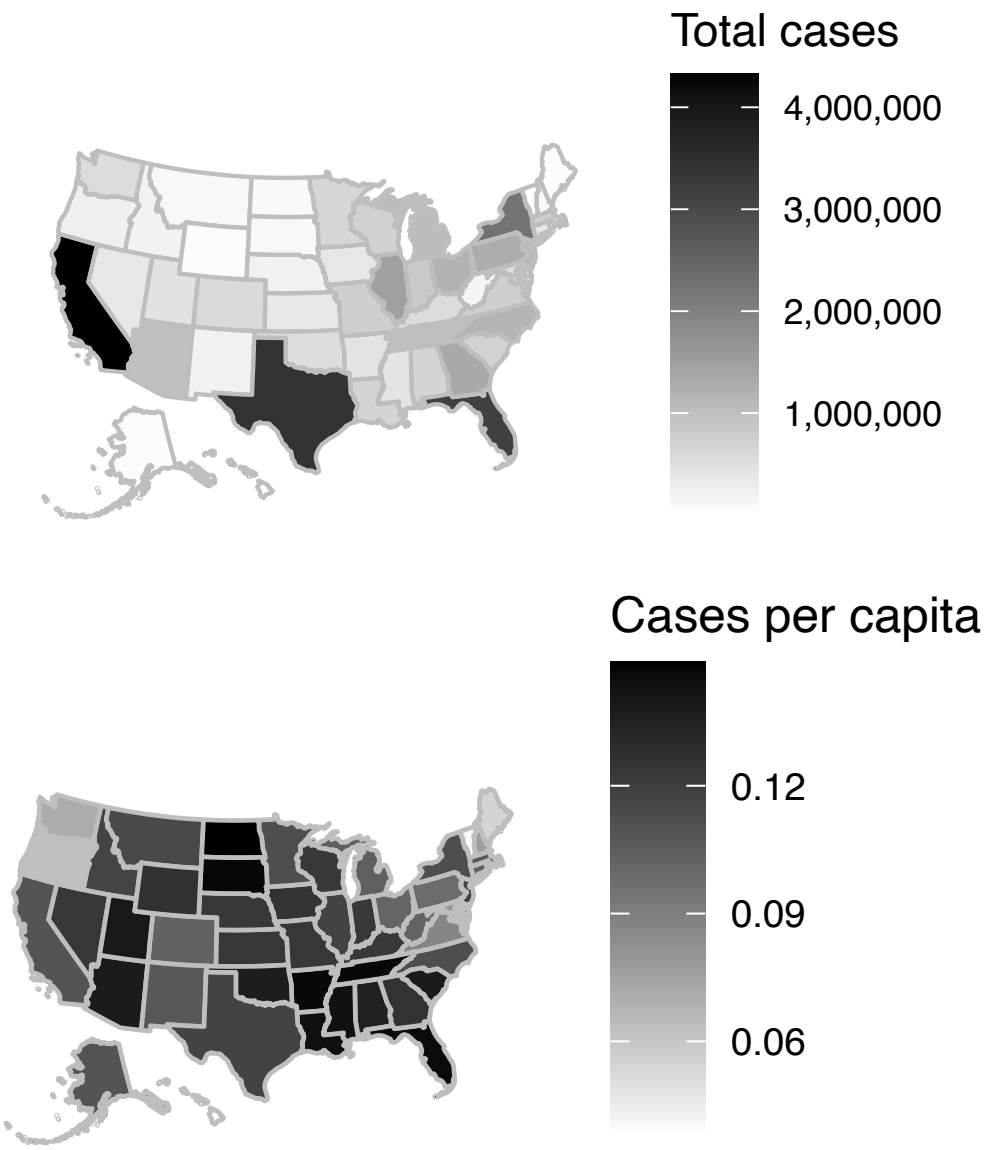

FIG 1. *(a) Statewide total cases recorded by the CDC from 1/23/20 to 08/25/21. (b) Statewide total cases per capita, recorded by the CDC from 1/23/20 to 08/25/21, using 2020 populations according to Census.gov .*

Cauchemez et al. 2006, Reinhart 2018). The model posits that $\lambda(t)$, the conditional rate at which points (confirmed cases) are expected to accumulate around time t, given information on all previous events, obeys

$$\lambda(t, x, y) = \mu + K(t) \sum_{i:t_i<t} g(t - t_i), \tag{1}$$

where μ is a constant background rate at which cases are thought to be entering the location externally, and g is a density function, i.e. nonnegative and integrating to 1, called the *triggering density* or *transmission time density*. $K(t)$ is called the *productivity*, indicating the expected number of points triggered directly by a point at time t, and is thus closely con-

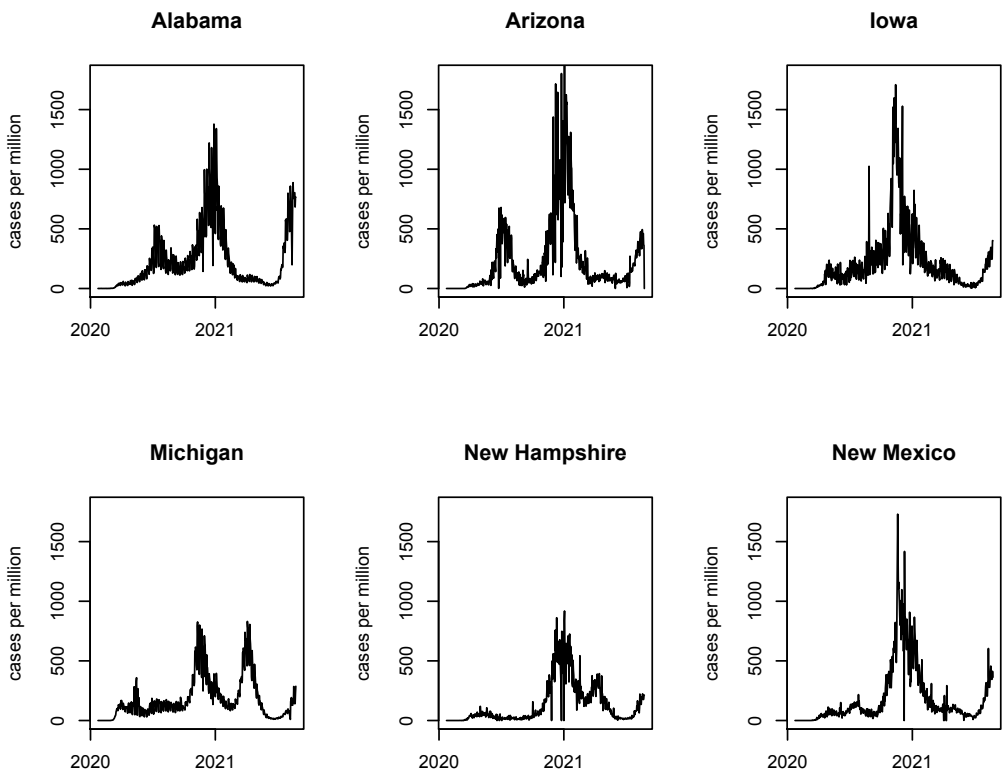

FIG 2. *(a) Recorded cases per million people, from 1/23/20 to 08/25/21, according to CDC statewide incidence data and 2020 population data from Census.gov .*

nected to the reproduction number (R_t) in compartmental models such as SEIR (Bertozzi et al. 2020, Kresin et al. 2021). Each background point associated with μ is expected to generate $K + K^2 + K^3 + ... = 1/(1-K) - 1 = K/(1-K)$ triggered points. As a result, in a Hawkes process with $0 < K < 1$, the expected fraction of background points is $1 - K$. Since the conditional rate uniquely characterizes the finite-dimensional distribution of any simple point process (Prop. 7.2.IV of Daley and Vere-Jones 2003), equation (1) fully specifies the model.

When the precise occurrence times of the individual cases are recorded in detail, the parameters in Hawkes processes are typically fit by maximum likelihood estimation (MLE), and the resulting estimates have desirable properties such as asymptotic unbiasedness, normality, consistency, and efficiency (Ogata 1978), even when minor covariates have been ignored (Schoenberg 2016). The triggering function can also be estimated non-parametrically (Marsan and Lengliné 2008), and some authors have also allowed the background rate μ to vary spatially and to be estimated nonparametrically, *e.g.* Zhuang et al. (2004) and Park et al. (2021). Bayesian methods can also be used to estimate parameters and quantify uncertainty in Hawkes process models (Rasmussen 2013, Mohler 2013).

Here, because the data consist of daily case counts rather than precise individual times within each day, the Hawkes model parameters are more efficiently fit by least squares, making use of the correspondence between Hawkes processes and autoregressive time series

derived in Kirchner (2016, 2017). That is, we find parameters minimizing the sum of squared differences between the observed and expected number of cases on each day, as specified as follows.

We fit both parametric and nonparametric forms of the model (1), where $K(t)$ is piecewise constant within each 16-day interval. For the non-parametric form of model (1), we assume $g(u)$ to be a step function with 16 steps, each of duration one day. This allows us to estimate a different step height for each day, so that the shape of the transmission time distribution can be determined by the data. Because the non-parametric estimates suggest a normal density might be a suitable approximation, we also consider a parametric form for g where

$$g(u) \sim N(\nu, \sigma^2) \tag{2}$$

is normally distributed with mean ν and standard deviation σ. For the parametric version of the model, the parameter vector θ for each state consists of μ, ν and σ, as well as 36 estimates of K, one for each 16-day interval. Thus 39 parameters are estimated. For the non-parametric form, μ and $K(t)$ are again estimated using 37 parameters, and in place of ν and σ, the 16 step heights corresponding to $g(t)$ are estimated, but since g is a density, $\sum_{i=1}^{16} g(i) = 1$, and thus $g(16) = 1 - \sum_{i=1}^{15} g(i)$, so only 15 free parameters govern g, and thus 52 parameters are to be estimated in total. For each such model, we find the value of the parameter vector, θ, minimizing

$$\sum_{t=1}^{T} \left\{ N(t) - [\mu + \sum_{i=1}^{16} K(t-i) g(i) N(t-i)] \right\}^2, \tag{3}$$

where $N(t)$ denotes the number of observed cases on day t, and $T = 36 \times 16 = 576$ days. Standard error estimates may be obtained via simulation and re-estimation. For each estimation, we used the Nelder-Mead optimization routine in the function *optim* in *R*, with a maximum number of iterations set to 100,000 and starting values of 1 for μ, σ, and $K(t)$, $1/16$ for each value of $g(u)$, and 10 for ν. Then for each estimation, the function *optim* was run again, with starting values equal to the ending values from the previous run, and again with a maximum number of iterations set to 100,000. When fitting Hawkes models for each of the 50 states, we fit the identical model in each state, but let all of the parameters differ from state to state.

4. Simulations. Simulations show that estimation of the transmission time distribution using (3) is very accurate. The processes are simulated as detailed in Section 3.3 of Reinhart (2018), or see Section 1 of the Supplementary Material (Schoenberg 2023). For each simulation, initially the background points are placed according to a homogeneous Poisson process of rate μ, and then each of these points triggers a random number of points according to a Poisson random variable with mean K, and these triggered points are distributed in time according to density g after the initial triggering point. These triggered points then trigger further points, and the process continues until no further points are generated within the given time window of 576 days.

Figure 3 shows the real and estimated transmission time densities for simulated Hawkes models with three very distinct normal densities with varying means and standard deviations, and in each simulation of 576 days the transmission time distribution is estimated quite

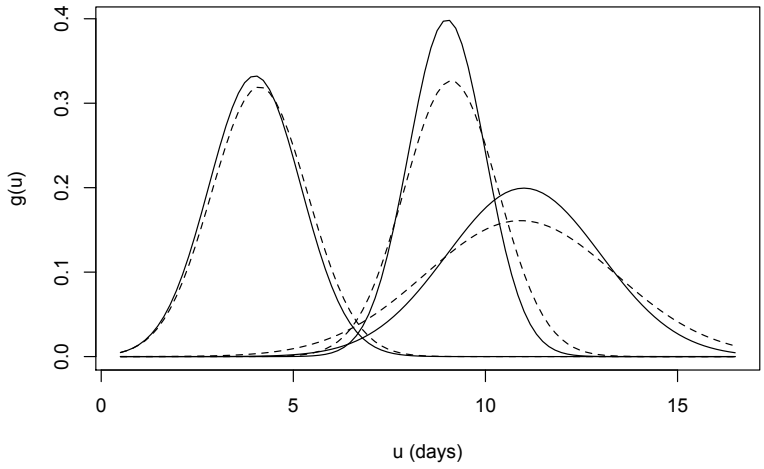

FIG 3. *True (solid curves) and estimated (dashed curves) transmission time densities for simulated Hawkes models. Each simulation is for 576 days, and for each simulation,* $\mu = 1$ *point/day and* $K = 0.95$ *points/day.* (ν, σ) *were (9 days, 1 day), (11 days, 2 days), and (4 days, 1.2 days).*

accurately. The parameters used in the first simulations were $\mu = 1$ point/day, $K = 0.95$ points/day, $\nu = 9$ days, and $\sigma = 1$ day, and for the other two simulations, μ and K were the same, and (ν, σ) = (11 days, 2 days) or (4 days, 1.2 days), respectively.

For the model with $\mu = 1$ point/day, $K = 0.95$ points/day, $\nu = 9$ days, and $\sigma = 1$ day, 50 simulations and re-estimations were performed, each consisting of 576 days, and the real and estimated transmission time densities are shown in Figure 4a. The root-mean-square (RMS) error in estimating ν and σ for the 50 simulations were 0.148 and 0.224, respectively. The least squares estimates are quite accurate for the duration corresponding to the SARS-CoV-2 data.

Figure 4b shows the real and estimated transmission time densities for simulated Hawkes processes with $N(9, 1)$ transmission times, but where in each simulation, 10% of the cases occurring on Saturdays and 20% of the cases occurring on Sundays are instead recorded on the following Monday. The resulting mean estimate of ν is 8.92 (the true value of ν is 9.0) and the mean estimate of σ is 1.56 (the true σ is 1.0). The RMS errors in estimating ν and σ for these simulations increased to 0.198 and 0.576, respectively. The least squares estimates appear to be slightly influenced by such recording errors.

5. Results. The Hawkes model fits the case count data very accurately, in all 50 states. Figure 5 shows the close approximation of the Hawkes model with nonparametrically estimated transmission density to the observed case counts throughout the observation period, for California, Pennsylvania, Florida, and Texas, for example, as well as the corresponding estimated productivities $K(t)$ for the Hawkes model fit to data from each of these four states. The RMS errors in daily case counts for the four states are 2699, 638, 1238, and 2729, respectively. The model with parametric (normal) transmission density had very similar RMS errors, corresponding to 2657, 670, 1190, and 2808 cases/day, for California, Pennsylvania,

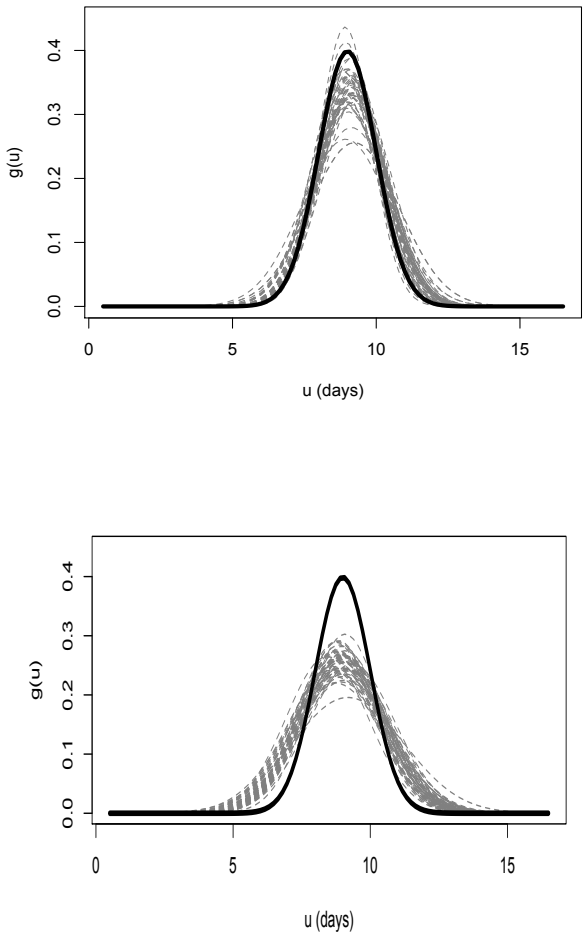

FIG 4. *True (solid black curve) and estimated (dashed grey curves) transmission time densities for 50 simulated Hawkes models (1). Each simulation is for 576 days, with parameters ($\mu = 1$ point/day, $K = 0.95$ points/day, $\nu = 9$ days, $\sigma = 1$ day). (a) Simulated Hawkes processes with no reporting errors. (b) Simulated Hawkes processes where 10% of cases on each Saturday and 20% of cases on each Sunday are recorded on the following Monday*

.

Florida, and Texas, respectively. For all 50 states, the RMSEs ranged from 22.1 for Vermont to 2729 for Texas, with a median RMS error of 355.3 corresponding to New Jersey, a mean RMS error of 493.7, a standard deviation of 539.4, and states with larger populations and higher case counts having correspondingly larger RMS errors.

Figure 6a shows the nonparametric transmission time density estimates for all 50 states, along with their mean. The peak at 7 days is clearly visible, and the normal density appears to be a reasonable approximation of the distribution, though the nonparametric estimates have some mass corresponding to transmission times of 1 day or 14 days. The estimated transmission period of 7 days may be characteristic of a weekly cycle in the data, though Figure 6b shows the confirmed case counts by weekday for each state and overall, and no clear weekly cycle is discerned. The state in Figure 6b with highly variable case counts is Kansas, where case counts were substantially higher on Wednesdays and Fridays, and lower on Tuesdays

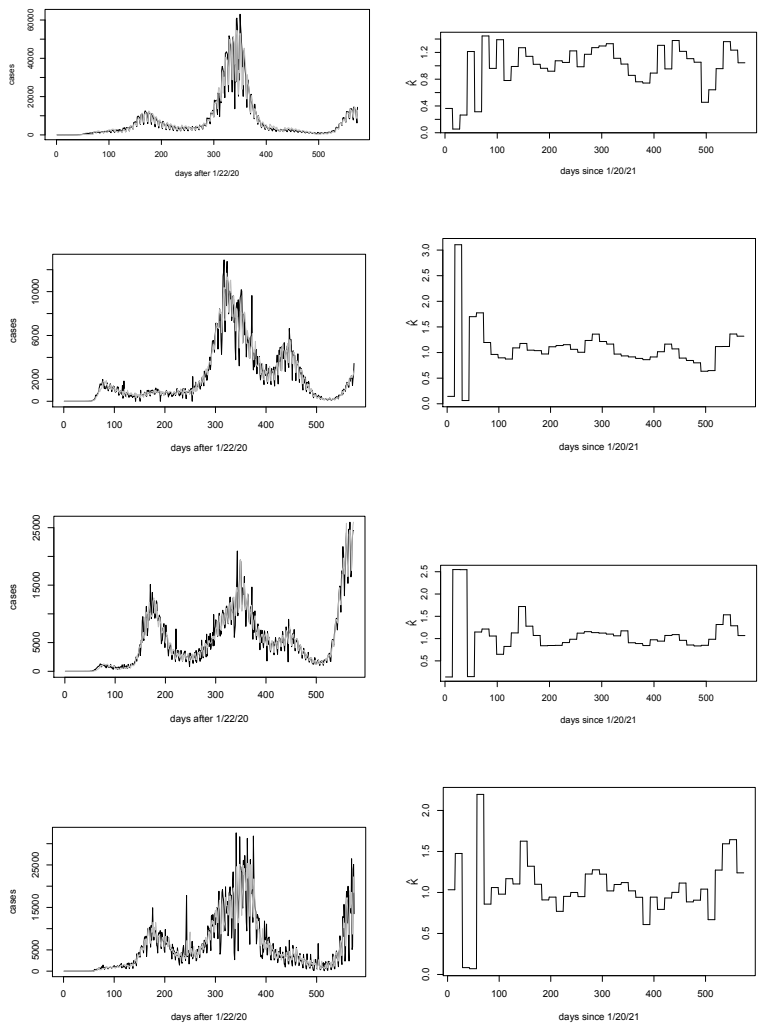

FIG 5. *Observed (black) and fitted (grey) confirmed case counts (left panels) and estimated productivities (right panels) for California (top row), Pennsylvania (2nd row), Florida (3rd row), and Texas (bottom row), fit using a Hawkes model (1) with nonparametrically estimated transmission density.*

and Thursdays, most likely due to reporting issues. Using the normal approximation for the transmission time distribution, the estimated transmission time densities for each of the 50 states are shown in Figure 6c, along with the mean. The outlying state with the unusually low estimated mean transmission time of 3.5 days is Ohio; the cause of this difference is unclear.

The corresponding estimates of ν and σ for the fitted normal transmission time densities for each of the 50 states are shown in Figure 7a. There is very close agreement among the estimates of ν from state to state, which the exceptions of Ohio, Virginia, Oklahoma and Kansas. For the other 46 states, the estimate of ν ranges from 6.51 to 7.22 days, with estimated standard errors between 0.10 and 0.21 days. The estimates of σ largely agree as well, with values ranging between 0.406 and 1.56 days, and with estimated standard errors between 0.15 and 0.27 days. A potential explanation for the four outlier states, particularly Kansas, could be the large number of days with zero confirmed cases, including numerous such days in late 2020 when case counts were high on most other days. Figure 7b shows the

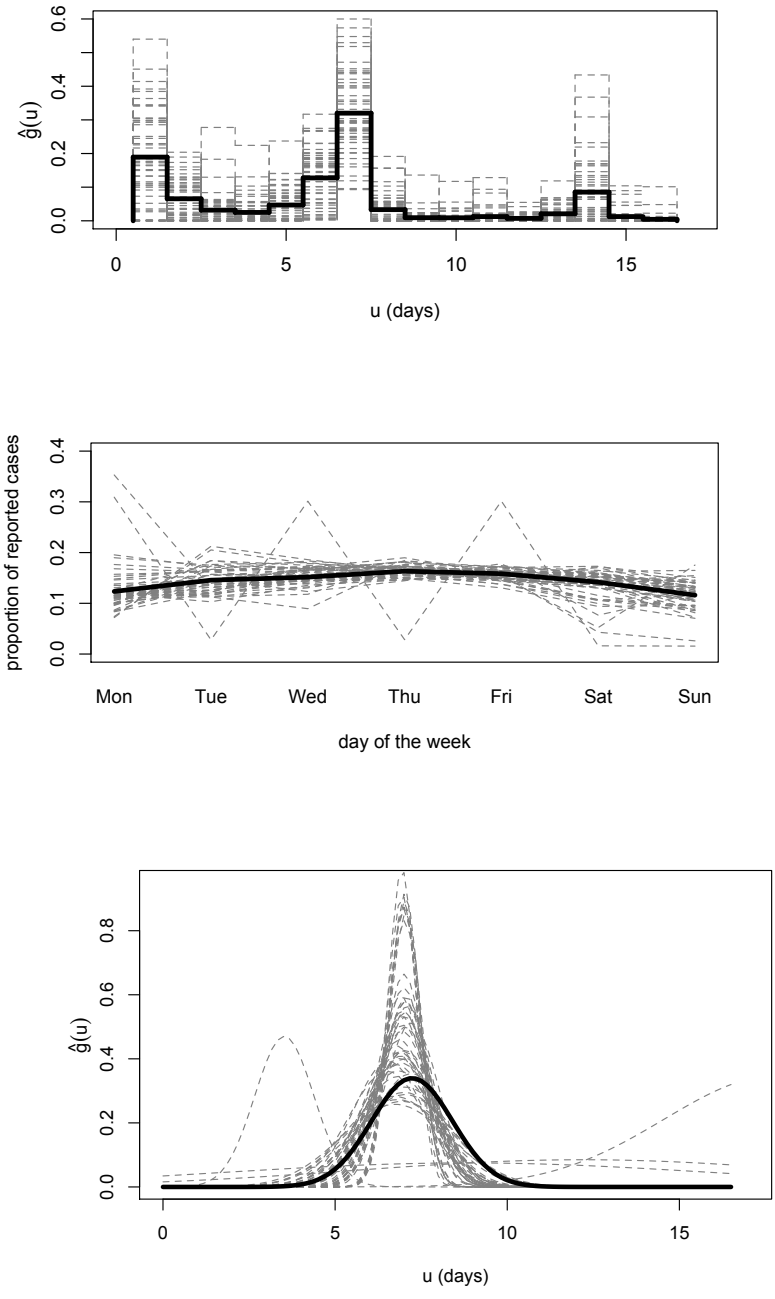

FIG 6. *(a) Nonparametric transmission time density estimates for all 50 states (thin grey curve), and mean of all 50 state estimates (thick black curve). (b) Proportion of case counts on each day of the week for all 50 states (thin grey curve), and mean of all 50 state proportions (thick black curve). (c) Estimated normal transmission time densities for all 50 states (thin grey curve), and mean of all 50 state estimates (thick black curve).*

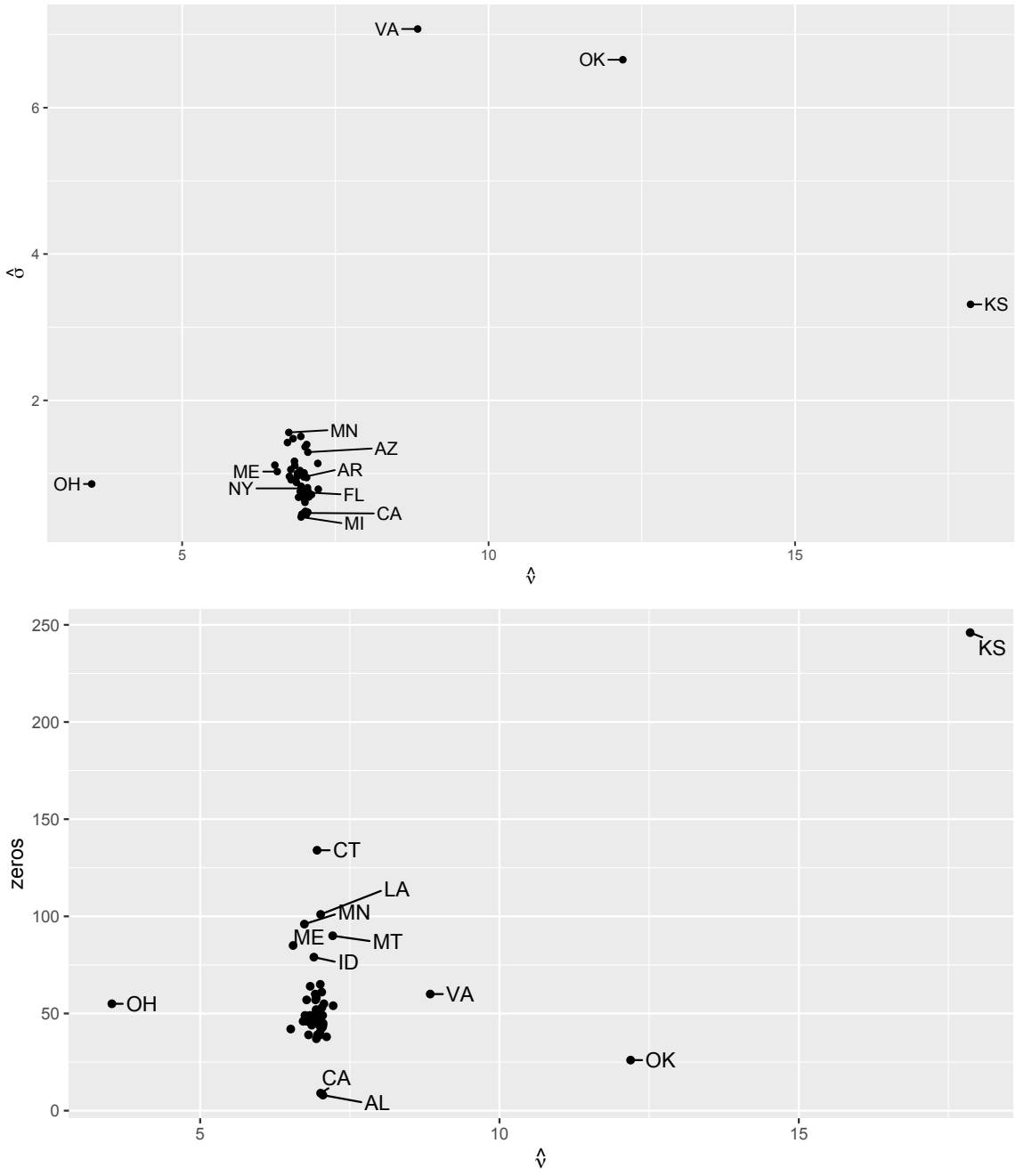

FIG 7. *(a) Estimates of the mean (ν) and standard deviation (σ) for estimated normal transmission time densities for fitted Hawkes models for all 50 states. (b) Estimates of the mean (ν) of the estimated normal transmission time density for fitted Hawkes models, for all 50 states, versus the number of days in each state with zero confirmed SARS-CoV-2 cases in the dataset.*

estimates of ν along with the number of days with zero confirmed case counts in the dataset, for each of the 50 states, and one sees that Kansas is a clear outlier.

6. Discussion. The results suggest that the transmission time distribution for SARS-CoV-2 is centered at 7 days, with a standard deviation of only approximately one day. This is

a somewhat more narrow density compared to prior reports based on case studies, including those by the CDC and WHO.

The density in the estimate of the triggering function at 14 days is likely due to harmonic aliasing (see e.g. Brillinger 1981). However, the estimated density at 1 day is more difficult to explain. One possibility might be contagion due to physical contact. While most of the contagion of SARS-CoV-2 may be attributable to aerial transmission from subjects with high viral loads via respiratory droplets, some transmission might be due to subjects spreading the virus very shortly after exposure (Lotfi et al. 2020). For example, if an individual's hand comes in contact with virus, then hand-to-hand exposure or hand-surface-hand exposure might be a means of transmission with very short transmission time (Lotfi et al. 2020). A more likely explanation may be that there is substantial autocorrelation between one day's case counts and the next for a variety of reasons, such as continuity in human behaviors, policies, and recording decisions, and the model attributes this autocorrelation to one-day transmission. An additional possible explanation is that some subjects may be infectious before their cases are reported, as the dates corresponding to recorded cases may differ from the actual dates of disease onset.

Data dumping and other data issues might also explain some of the results. Indeed, the high rate of estimated transmissions at 7 days might at least partially be the result of regular weekly trends in reporting of cases. If cases discovered on weekends are more likely to be reported on Mondays, for instance, then the higher reported case loads on Mondays may be expected to yield higher estimated transmission time density at one week. However, as shown in Figure 4b, the transmission time estimates are not extremely sensitive to such errors, and the numbers of confirmed cases hardly vary by weekday in most states as well as overall as shown in Figure 6b. In addition, the results were remarkably consistent across states, with the estimated density peaking at 7 days from state to state.

It is important to remember, however, that many cases of SARS-CoV-2 are likely missing from the state catalogs and the CDC database. In particular, while unreported asymptomatic cases likely were common throughout the pandemic, under-reporting was most likely more prevalent earlier in the pandemic, when tests were more scarce. The results here, which pertain to the transmission time distribution, are most likely minimally affected by this variable under-reporting, unless the under-reporting is occurring in some systematic fashion, and there is presently little evidence of this, other than perhaps the weekly trends referred to above. As discussed in Kresin et al. (2021), the difficulty in estimating the percentage of asymptomatic cases appears to introduce more error into SEIR models than Hawkes models because of the high sensitivity to this parameter in SEIR models, as opposed to the Hawkes model which essentially bypasses the issue of estimating the number of asymptomatic cases by instead focusing on the rate of recorded cases as a function of prior recorded cases, essentially like an autoregressive process.

The assumption of a constant value of μ, for each of the 50 states, may be violated since any particular state may have rates of immigration of the virus that vary over time, particularly as stay-at-home orders changed over time. Again, however, such errors in the estimation of the background rate would be unlikely to result in substantial errors in the estimation of the transmission time.

As an alternative to the method employed here of fitting a simple Hawkes model separately to data from each of the 50 states would be to fit a single large, multi-variate Hawkes model

to data from all 50 states simultaneously, and where transmission from place to place is permitted. This type of modeling was explored for example in Yuan et al. (2021). Problems of non-identifiability and especially multicollinearity can be very severe when fitting such large models with many parameters, however, leading to estimates with high variance and poor forecasting performance. As noted in Yuan et al. (2021), when fitting a multivariate model to Covid-19 data from different cities, the data are often so highly correlated that the estimated cross-productivities are often highly unrealistic. Addressing this issue is an important subject for future statistical work.

A very important topic for future research is the reliable estimation of standard errors and other measures of uncertainty for Hawkes model parameters or forecasts, especially for Hawkes models with variable productivity estimated by least squares, and with nonparametrically estimated triggering functions. Currently the best available methods seem to be simulation-based, but for variable-productivity Hawkes processes with triggering functions estimated nonparametrically, simulation of the estimated model followed by subsequent re-estimation of the triggering function and varying productivities seems to result in somewhat underestimated uncertainties (e.g. Schoenberg 2022). How best to inflate such uncertainty estimates is an important subject for future research.

SUPPLEMENTARY MATERIAL

Code for Data Analysis and Simulations
Zip file containing R code used for the data analysis, simulations, and construction of Figures 1-7.

Acknowledgments. This material is based upon work supported by the National Science Foundation under grant number DMS-2124433. We also thank the CDC for providing the data for our analysis. Computations were performed in *R*. The code for our simulations and computations is in Section 1 of the Supplementary Material (Schoenberg 2023).

Funding. The author was supported by NSF Grant DMS-2124433

References

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, and Kyle T (2021). Estimated SARS-CoV-2 Seroprevalence in the US as of September 2020. *JAMA Intern Med.* 181(4):450-460.

Bertozzi AL, Franco E, Mohler G, Short MB, and Sledge D (2020). The challenges of modeling and forecasting the spread of COVID-19. *Proceedings of the National Academy of Sciences* 117(29), 16732-16738.

Bertsimas D (2020). MIT Covidanalytics, May 2020. covidanalytics.io [Online; accessed 24-May-2020].

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.

Brillinger DR (1981). Time Series: Data Analysis and Theory. Holden-Day, San Francisco.

Cauchemez S, Boelle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, Hedley AJ, Anderson RM, and Valleron AJ (2006). Real-time estimates in early detection of SARS. Emerging infectious diseases, 12(1):110.

Centers for Disease Control and Prevention (CDC) (2021a). https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html , last accessed 9/14/21.

Centers for Disease Control and Prevention (CDC) (2021b). https://www.cdc.gov/coronavirus/2019-ncov/your-health/quarantine-isolation.html , last accessed 9/14/21.

Centers for Disease Control and Prevention (CDC) (2021c). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html , last accessed 9/14/21.

Centers for Disease Control and Prevention (CDC) (2021d). https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html , last accessed 9/14/21.

Centers for Disease Control and Prevention (CDC) (2021e). https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm , last accessed 9/14/21.

Centers for Disease Control and Prevention (CDC) (2021f). https://covid.cdc.gov/covid-data-tracker , last accessed 9/14/21.

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

Clements, R.A., Schoenberg, F.P., and Schorlemmer, D., 2011. Residual analysis for space-time point processes with applications to earthquake forecast models in *Annals of Applied Statistics* **5**(4), 2549–2571.

Clements RA, Schoenberg FP, Veen A (2013). Evaluation of space-time point process models using super-thinning. *Environmetrics* 23(7), 606-616.

Daley DJ and Vere-Jones D (2003). An Introduction to the Theory of Point Processes: Elementary Theory of Point Processes. Springer, 2003.

Dye C and Gay N (2003). Modeling the SARS epidemic. Science, 300(5627):1884-1885.

Farrington C, Kanaan M, and Gay N (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics, 4(2):279-295.

Gordon JS, Clements RA, Schoenberg FP, and Schorlemmer D (2015). Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. Spatial Statistics, 14:133-150.

Guan WJ, Ni ZY, Hu Y, et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med.* 382, 1708-1720.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *J. Roy. Statist. Soc.*, **B33**, 438-443.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 395 (10223), 497-506.

Institute for Health Metrics and Evaluation (IHME) (2020). IHME Covid-19 predictions, May 2020. covid19.healthdata.org [Online; accessed 24-May-2020].

Jewell NP, Lewnard JA, and Jewell BL (2020). Caution warranted: using the Institute for Health Metrics and Evaluation model for predicting the course of the Covid-19 pandemic. *Annals of Internal Medicine* 173(3), 226-227.

Kelly, J.D., R.J. Harrigan, J. Park, N.A. Hoff, S.D. Lee, R. Wannier, B. Selo, M. Mossoko, B. Njokolo, E. Okitolonda-Wemakoy, P. Mbala-Kingebeni, G.W. Rutherford, T.B. Smith, S. Ahuka-Mundeke, J.J. Muyembe-Tamfum, A.W. Rimoin, and F.P. Schoenberg (2019). Real-time predictions of the 2018-2019 Ebola virus disease outbreak in the Democratic Republic of Congo using Hawkes point process models. *Epidemics* 28, 100354.

Kirchner, M (2016). Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications* 26(8), 2494-2525.

Kirchner M (2017). An estimation procedure for the Hawkes process. *Quant Financ.* 17(4):571-595.

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

Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Ann Intern Med.* 172(9):577-582.

Lekone PE and Finkenstadt BF (2006). Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. Biometrics, 62(4):1170-1177.

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med.* 382, 1199-207.

Los Alamos National Laboratory (LANL) (2020). Covid-19 confirmed and forecasted case data, May 2020. covid-19.bsvgateway.org [Online; accessed 24-May-2020].

Lotfi M, Hamblin MR, and Rezaei M. (2020). COVID-19: Transmission, prevention, and potential therapeutic opportunities. *Clin. Chim. Acta* 508, 254-266.

Marsan D, Lengliné O (2008). Extending earthquakes' reach through cascading. *Science* 319(5866), 1076-1079.

Meyers LA (2007). Contact network epidemiology: bond percolation applied to infectious disease prediction and control. *Bull. Amer. Math. Soc.* 44(1), 63-86.

Mohler G (2013). Modeling and estimation of multi-source clustering in crime and security data. *Annals of Applied Statistics* 7(3), 1525-1539.

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

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

Ogata, Y. (1988), 'Statistical models for earthquake occurrence and residual analysis for point processes', *J. Amer. Statist. Assoc.*, **83**, 9-27.

Park J., Chaffee A., Harrigan R., and Schoenberg F.P. (2020). A non-parametric Hawkes model of the spread of Ebola in West Africa. *J. Appl. Stat.*, 1-27.

Park J, Schoenberg FP, Bertozzi AL, and Brantingham PJ (2021). Investigating clustering and violence interruption in gang-related violent crime data using spatial-temporal point processes with covariates. JASA, in press.

Rasmussen JG (2013). Bayesian inference for Hawkes processes. Methodology and Computing in Applied Probability, 15(3):623-642.

Reinhart A (2018). A review of self-exciting spatio-temporal point processes and their applications. Statistical Science, 33(3):299-318.

Rizoiu MA, Mishra S, Kong Q, Carman M, and Xie L. (2018). Sir-Hawkes: linking epidemic models and Hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference*, pp. 419?428.

Schoenberg, F.P. (2016). A note on the consistent estimation of spatial-temporal point process parameters. *Statistica Sinica*, 26, 861-789.

Schoenberg F. (2022). Nonparametric estimation of variable productivity Hawkes processes. *Environmetrics* 33(6), e2747.

Schoenberg F. (2023). Supplement to "Estimating Covid-19 transmission time using Hawkes point processes." DOI: 10.1214/

Schoenberg, F.P., Hoffmann, M., and Harrigan, R. (2019). A recursive point process model for infectious diseases. *AISM* 71(5), 1271-1287.

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 (2018). The Collaboratory for the Study of Earthquake Predictability: achievements and priorities. Seismological Research Letters, 89(4):1305-1313.

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

Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y, Zhao Y, Li Y, Wang X, Peng Z (2020). Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* 323(11), 1061-1069. doi:10.1001/jama.2020.1585.

Wermer E and Stein J (2020). Trump administration pushing to block new money for testing, tracing and CDC in upcoming coronavirus relief bill. *Washington Post*, 07/18/20, https://www.washingtonpost.com/us-policy/2020/07/18/white-house-testing-budget-cdc-coronavirus.

World Health Organization (WHO) (2021a). https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions , last accessed 9/14/21.

World Health Organization (WHO) (2021b). https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions , last accessed 9/14/21.

Xu RH, He JF, Evans MR, Peng GW, Field HE, Yu DW, Lee CK, Luo HM, Lin WS, Lin P, Li LH, Liang WJ, Lin JY, and Schur A (2004). Epidemiological clues to SARS origin in China. *Emerg. Infect. Dis.* 10(6), 1030-1037.

Yang AS (2019). Modeling the Transmission Dynamics of Pertussis Using Recursive Point Process and SEIR model. PhD thesis, UCLA.

Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, Wu Y, Zhang L, Yu Z, Fang M, Yu T, Wang Y, Pan S, Zou X, Yuan S, Shang Y (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med.* 8(5), 475-481.

You C, Deng Y, Hu W, Sun J, Lin Q, Zhou F, Pang CH, Zhang Y, Chen Z, and Zhou XH (2020). Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *International Journal of Hygiene and Environmental Health*, 113555.

Yuan B, Schoenberg FP, and Bertozzi AL (2021). Fast estimation of multivariate spatiotemporal Hawkes processes and network reconstruction . AISM 73 (6), 1127-1152.

Zechar, J. D., Schorlemmer, D., Werner, M.J., Gerstenberger, M.C., Rhoades, D.A., and Jordan, T.H. (2013). Regional Earthquake Likelihood Models I: First-order results. *Bull. Seismol. Soc. Am.* , 103, 787-798.

Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Hui L, Wu X, Xu J, Tu S, Zhang Y, Chen H, Cao B (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 395, 1054-1062. doi.org/10.1016/S0140-6736(20)30566-3.

Zhuang J, Ogata Y, and Vere-Jones D (2004). Analyzing earthquake clustering features by using stochastic reconstruction. Journal of Geophysical Research: Solid Earth, 109(B5).