Epidemics 51 (2025) 100824

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Detection of surges of SARS-Cov-2 using nonparametric Hawkes models

Sophie Phillips [a], George Mohler [b], Frederic Schoenberg [a,*]

[a] Department of Statistics, 8142 Math-Science Building, UCLA, Los Angeles, CA 90095-1554, USA
[b] Department of Computer Science, Boston College, 245 Beacon Street, Boston, MA 02116, USA

ARTICLE INFO

Dataset link: https://github.com/CSSEGISandData/COVID-19

Keywords:
Covid-19
Effective reproduction rate
Epidemiological forecasting
Forecasting
Hawkes model
Point process

ABSTRACT

Hawkes point process models have been shown to forecast the number of daily new cases of epidemic diseases, including SARS-CoV-2 (Covid-19), with high accuracy. Here, we explore how accurately Hawkes models forecast surges of Covid-19 in the United States. We use Hawkes models to estimate the effective reproduction rate R_t and transmission density parameters for Covid-19 case counts in each of the 50 United States, then forecast R_t in future weeks with simple exponential smoothing. A classifier based on $R_t > x$ is applied to predict upcoming surges in cases each week from August 2020 to December 2021, using only data available up to that week. At false alarm rates below 5%, the forecasts based on R_t are correct more often than forecasts based on smoothing the raw case count data, achieving a maximum accuracy of 90% with $R_t > 1.39$. The optimal decision boundary uses a combination of R_t and observed data.

1. Introduction

Hawkes point process models and their extensions such as the epidemic-type aftershock sequence (ETAS) model (Ogata, 1988, 1998), HawkesN (Rizoiu et al., 2018; Bertozzi et al., 2020; Chiang et al., 2022), and the recursive model (Schoenberg et al., 2019) have proven to be useful in modeling the spread of a variety of epidemic diseases such as Ebola (Harrigan et al., 2019; Park et al., 2022), Chlamydia (Schoenberg, 2022), SARS (Wallinga and Teunis, 2004; Cauchemez et al., 2006), measles (Farrington et al., 2003), Meningococcal disease (Meyer et al., 2012), Rocky Mountain Spotted Fever (Schoenberg et al., 2019), and SARS-CoV-2 (Covid-19) (Mohler et al., 2021; Schoenberg, 2023; Phillips and Schoenberg, 2024). Kresin et al. (2022) compared the accuracy of Hawkes models to that of compartmental models, such as SEIR, across various infectious diseases, and found that Hawkes models offered substantially higher accuracy than compartmental models, with errors in forecast case counts approximately 20%–30% smaller on average than compartmental models in most cases.

Here, we explore how accurately Hawkes models can forecast surges of Covid-19 in the United States. One way to evaluate the model's forecasting efficacy is to consider its accuracy in estimating the effective reproduction rate, which is an important and widely used indicator for the transmission intensity of an epidemic and an early warning signal for disease emergence (Anderson, 1991; Southall et al., 2020). This parameter, denoted R_t, represents the mean number of secondary cases generated by a new infection at time t (Cori et al., 2013). In theory, R_t is as an early warning statistic for outbreaks, with values above 1 indicating an emerging disease (Southall et al., 2020). Accurate estimates of R_t can be used to inform surveillance efforts and assess the impact of interventions (Eichner and Dietz, 2003; Ferguson et al., 2006; Domenech de Cellès et al., 2018; Bertozzi et al., 2020; Chiang et al., 2022).

R_t can be estimated retrospectively or in real time. Wallinga and Teunis (2004) use an expectation maximization (EM) approach to reconstruct the transmission chains and estimate the number of secondary cases per individual. Similarly, Mohler, Schoenberg, and Sledge (2021) apply an EM algorithm to estimate R_t over predefined intervals while simultaneously estimating transmission parameters. These methods are only suitable for retrospective analysis as the estimates of R_t at a given time rely on subsequent case records, although modifications for real-time estimation have been proposed (Cauchemez et al., 2006). Alternatively, Cori et al. (2013) use Bayesian inference with a Gamma prior on R_t, resulting in a Gamma posterior under a Poisson likelihood for incident cases. Other methods assume R_t changes at estimated or known change points (Browning et al., 2021; Chen et al., 2021), including a state-space method for detecting change points in R_t based on a discrete-time Hawkes model (Koyama et al., 2021). In contrast to these methods, we avoid distributional assumptions on the parameter, allowing R(t) to vary freely over time, and use a direct least-squares estimation method. We also present an alternating least-squares procedure for estimating the transmission parameters.

In this paper, we examine the use of the estimated reproduction rate for forecasting outbreaks in incidence, using transmission parameters

* Corresponding author.
E-mail address: frederic@stat.UCLA.edu (F. Schoenberg).

https://doi.org/10.1016/j.epidem.2025.100824
Received 28 October 2024; Received in revised form 17 March 2025; Accepted 25 March 2025
Available online 12 April 2025

estimated from a discrete Hawkes model with state-level Covid-19 incidence data from the first six months of 2020. Case count data are incrementally added week by week and R_t is sequentially re-estimated. We assess the forecast accuracy by observing how frequently the estimated reproduction rate exceeding a threshold precedes surges in cases, and compare with forecasts based solely on prior case counts.

The structure of the paper is outlined as follows. Statewide data on Covid-19 case counts data are described in Section 2, followed by a description of the model and estimation methods in Section 3. These methods are applied to estimate the transmission parameters and reproduction rate for all 50 states, and the resulting estimates are further analyzed to see how well they could have forecast surges in incidence. Results are detailed in Section 4, and a discussion is given in Section 5.

2. Data description

Daily counts of new Covid-19 cases were obtained for each state from the Johns Hopkins Covid-19 Data Repository which sourced case data from the CDC and state public health departments (Dong et al., 2020; JHUCSSE, 2023). The case counts include both cases confirmed by a positive PCR test and probable cases defined by a combination of antigen testing and epidemiological criteria (CDC, 2021). Positive tests are dated either according to the date of report or date of sample collection (CDC, 2021). Daily case counts for each state were obtained for the period 2/1/2020 to 12/31/2021, covering a total of 100 weeks. We ended the study period at the end of 2021 as many states shifted to less frequent reporting schedules in 2022, causing irregularities in the case count data (Dong et al., 2022). We model here only confirmed and suspected cases, though many Covid-19 infections are asymptomatic or unrecorded (Ma et al., 2021). The data used are publicly available from https://github.com/CSSEGISandData/COVID-19.

During the study period, most states experienced two major surges in reported cases beginning in November 2020 and 2021 and each lasting until the following March, although the timing, duration, and magnitude of these surges varied substantially between states. The reported case counts for California, Texas, Vermont, and Hawaii are shown in Fig. 1 as examples. California, Texas, and Florida experienced the highest total number of new cases during the study period, with 5.3, 4.5, and 3.9 million cases respectively. North Dakota and Alaska experienced the highest per capita incidence during this period, while Hawaii and Vermont experienced both the lowest per capita incidence as well as the least total incidence, with approximately 99,900 and 60,200 total cases, respectively, during this study period. Hawaii and Vermont experienced modest increases in reported case counts during 2020, whereas California and Texas had significant increases in both the Summer and Winter of 2020. Sudden spikes in case count records, such as those in Texas in early 2021 and in Vermont in late 2021, appear in several states throughout the study period.

3. Methods

3.1. Discrete Hawkes model

The Hawkes self-exciting point process (Hawkes, 1971) models the rate of infections $\lambda(t)$ conditional on previous infection times t_i as

$$\lambda(t) = \mu + \sum_{t_i<t} R(t_i)g(t-t_i) \tag{1}$$

The transmission density $g(t)$ and productivity function $R(t)$ are directly analogous to those in epidemiological models, and accommodate a variety of parametric and non-parametric forms. The parameter μ represents the rate of new cases immigrating into the location of interest, and is typically assumed to remain constant over time, though it can also be modeled as a function of time and/or space.

In epidemiological applications, Hawkes models have demonstrated superior forecasting accuracy and flexibility relative to traditional compartmental models (Bertozzi et al., 2020; Mohler et al., 2021; Chiang et al., 2022; Kresin et al., 2022). Following Phillips and Schoenberg (2024), we use a discrete approximation of the continuous time Hawkes model (1) to model the case count rather than rate, as cases are reported as daily counts rather than precise event times. Kirchner (2016) shows this binned form of Hawkes model corresponds to an autoregressive model. We fit the model by finding parameters that minimize the squared error between the observed and modeled daily case counts, as in Schoenberg (2023).

3.2. Non-parametric estimation of R(t)

We fit the reproduction rate non-parametrically as a step function over weekly intervals, thus estimating a total of 100 parameters for each state over the study period. Each estimate represents the expected number of infections directly transmitted by each case reported that week. In this section, the transmission time density and baseline rate μ are assumed to be known. In the next section, we present a method for jointly estimating the transmission mean, baseline, and reproduction rate $R(t)$. $R(t)$ is assumed constant within each non-overlapping τ-length interval. The window length τ should be large enough to reduce the number of intervals with zero cases, which inflate uncertainty estimates, but not so large that it delays the detection of changes in $R(t)$. We find that weekly intervals provide reasonably smooth estimates for the COVID-19 data.

The least-squares estimates for $R = \left(R_1 \dots R_{T/\tau}\right)$ can be obtained analytically using the method proposed for continuous-time Hawkes processes in Schoenberg (2022) and for discrete case counts in Phillips and Schoenberg (2024). We estimate the final estimate $R_{T/\tau}$ with exponential smoothing. Below, we describe the method for the general case of estimating R(t) over T days using a window length τ, yielding T/τ parameters to estimate.

We seek to minimize the error in estimating the case count $N(t)$ for day t:

$$\sum_{t=2}^{T}[N(t) - \mu - \sum_{s=1}^{t-1} R(s)g(t-s)N(s)]^2 \tag{2}$$

Let N be the vector of observed daily counts excluding the first day, and G be the $(T-1)\times(T/\tau-1)$ matrix with entries $G_{ij} = \sum_{s=(j-1)\tau+1}^{j\tau} g(i-s)N(s)$. G_{ij} represents the total expected number of new cases on day i directly infected from cases occurring in window j. With this notation, (2) can be expressed as:

$$\hat{N} = G\vec{R} \tag{3}$$

A column of 1s can be appended to G to estimate the baseline intensity μ as an intercept. When G is invertible, the estimates $\vec{R}$ can be computed quickly and directly using standard least-squares solvers. However, the matrix G is generally sparse and ill-conditioned, with a banded structure determined by τ and the length of the transmission interval. When $\tau = 1$, G is lower triangular. While increasing τ can reduce sparsity by incorporating more cases into each interval, this effect is limited as $g(t)$ is typically only positive over relatively short intervals. As in Phillips and Schoenberg (2024), to avoid issues with matrix singularity and improve the stability of estimates, we regularize the system in (3) by adding a penalty on the solution norm:

$$\hat{\vec{R}} = \min_{R_1\dots R_{T/\tau-1}} \|\vec{N} - \mu - G\vec{R}\|_2^2 + \rho\|\vec{R}\|_2^2 \tag{4}$$

The regularization parameter ρ controls the degree of smoothing. We incrementally increase ρ until all estimates are non-negative, searching over values of $10^{-k}\sigma_1(G)$, where σ_1 is the largest singular value of G. The standard errors for the regularized least-squares estimates can be calculated as:

$$\mathrm{Var}(\hat{R}) = s^2(G^TG + \rho I)^{-1}, \tag{5}$$

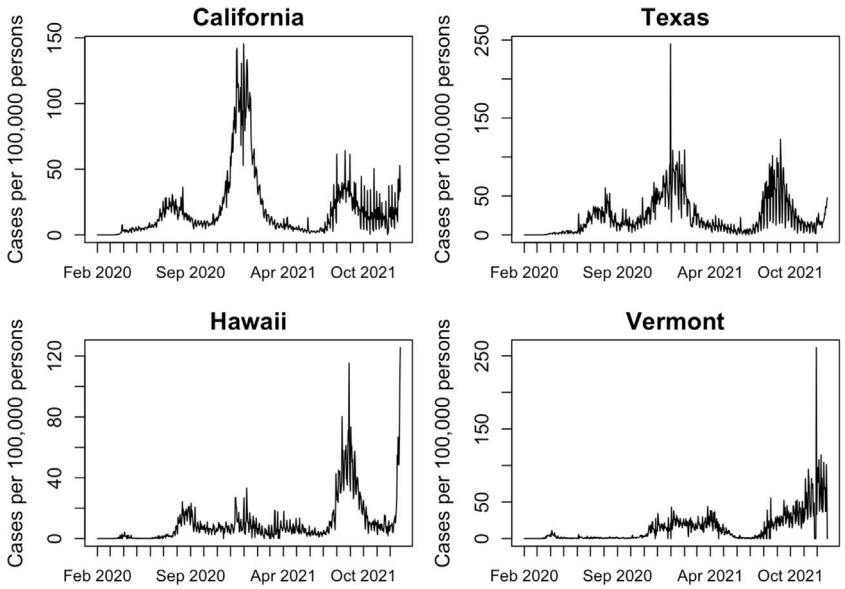

Fig. 1. Daily reported cases per 100,000 persons in California, Texas, Hawaii, and Vermont. Population estimates are from the 2020 decennial census.

$$s^2 = \frac{1}{T-d}\|\vec{N}(t) - \hat{\vec{N}}(t)\|_2^2,$$

where $d = T/\tau$. The covariance matrix (5) can be calculated efficiently using the Cholesky factorization of $G^T G + \rho I$.

In order to emulate real-time forecasting, we apply simple exponential smoothing to forecast $R_{T/\tau}$ from the least-squares estimates $R_1 \dots R_{T/\tau-1}$ calculated using (4). Exponential smoothing effectively captures short-term trends in time series without imposing strong assumptions about the underlying data structure, which accommodates our non-parametric estimates of $R(t)$. This method forecasts values by applying a weighted average of previous estimates, with weights that decay exponentially at a rate of α. We use $\alpha = 0.5$ here. The variance of the forecast value $R_{T/\tau}$ is calculated based on the residual variance from earlier forecasts (Brown, 1963).

For comparison, we present daily estimates of $R(t)$ from EpiNow2 (Abbott et al., 2024), the model used by the Centers for Disease Control's outbreak analytics team (CDC, 2024). This model is based on the Bayesian method for estimating R(t) proposed in Cori et al. (2013), but assumes a zero-mean Gaussian Process prior on the first differences, $\log R_t - \log R_{t-1}$. We set the prior for the initial reproduction number to have mean 2 and standard deviation 1. For both methods of estimating $R(t)$, we use the same transmission parameters which are estimated in the following section.

3.3. Estimation of transmission density parameters

Transmission parameters are estimated for each state using data from 2/1/2020 to 7/31/2020. Since the transmission time density for statewide Covid-19 data was shown in Schoenberg (2023) to be approximately normal, we model the transmission time density parametrically using a truncated normal distribution: $g \sim N_{\geq 0}(\nu, \sigma)$, and estimate the mean ν for each state. To avoid overfitting the training period, we fix the standard deviation σ at 1 based on the range of 0.4 to 1.56 estimated by Schoenberg (2023) across all states.

The reproduction rate $R(t)$ is fit simultaneously with the transmission mean and baseline rate as follows. Starting from initial estimates $\vec{R}^{(0)}$ and $\mu^{(0)}$, a non-linear optimizer determines ν that minimizes the squared error in daily cases. $\vec{R}$ and μ are then calculated analytically based on (4), with the matrix G formed using the current estimates of ν and augmented to include an intercept for estimation of μ. The final estimate of $\vec{R}$ is estimated with exponential smoothing as detailed in the previous section. We iterate between these steps until the estimates of ν, μ, and $\vec{R}$ converge. This algorithm is detailed in Algorithm 1. For simplicity, we initialize $\vec{R}^{(0)}$ as a constant vector of 0.5 and $\mu^{(0)} = \bar{N}(t)$, and use the BFGS routine in R's optim. We set the convergence threshold to $\epsilon = 10^{-2}$ and the maximum number of iterations to 100.

3.4. Forecasting surges using R_t

To mimic estimation in real-time, we sequentially incorporate case data week by week from 8/1/2020 to 12/25/2021, re-estimating R(t) using only data up to the latest available date for a total of 72 forecast periods. For any time t, the estimate of $R(t)$ is used to forecast upcoming surges in cases just after time t.

Early warning signals are often evaluated with synthetic data, which can struggle to capture the biases and noise in real data (Unkel et al., 2012; Southall et al., 2020). Methods for validating signals with real data include checking whether signals coincide with rising incidence or public health interventions (Brett and Rohani, 2020; Liu et al., 2021), defining outbreaks based on a window around the peak incidence date (Cowling et al., 2006; Son et al., 2020), or estimating the inflection points of the series with smooth curves (Kaur et al., 2020; O'Brien and Clements, 2021). Here, we define a surge as the times when the smoothed weekly case count increases by at least 10% for more than three consecutive weeks. This criterion effectively captures major outbreaks. Fig. 2 highlights the surge periods for a sample of states, with data for all 50 states shown in Figure S1. We use $R_t > x$, where x is some threshold, as a classifier to predict whether a surge, as defined above, will occur either one or two weeks after time t. This method is compared to an "incidence-based" classifier based on the percent change in weekly incidence: $\frac{\sum_{t\in w} N(t) - N(t-7)}{\sum_{t\in w} N(t-7)} > y$, where $N(t)$ is the case count at time t and w represents a given week. It is also compared to a "random guesser" null model that simply predicts surge times completely at random. We evaluate the performance of each classifier using a receiver operator characteristic (ROC) curve which reports the true positive and false rate across a range of thresholds x, y.

A linear support vector machine (SVM) is used to determine the line that maximally separates weeks during surges from non-surge weeks based on the previous week's predicted R_t and change in incidence (Boser et al., 1992). The SVM is trained to predict whether the 72 forecast periods in each state are during a surge using the observed and estimated measures from the previous week.

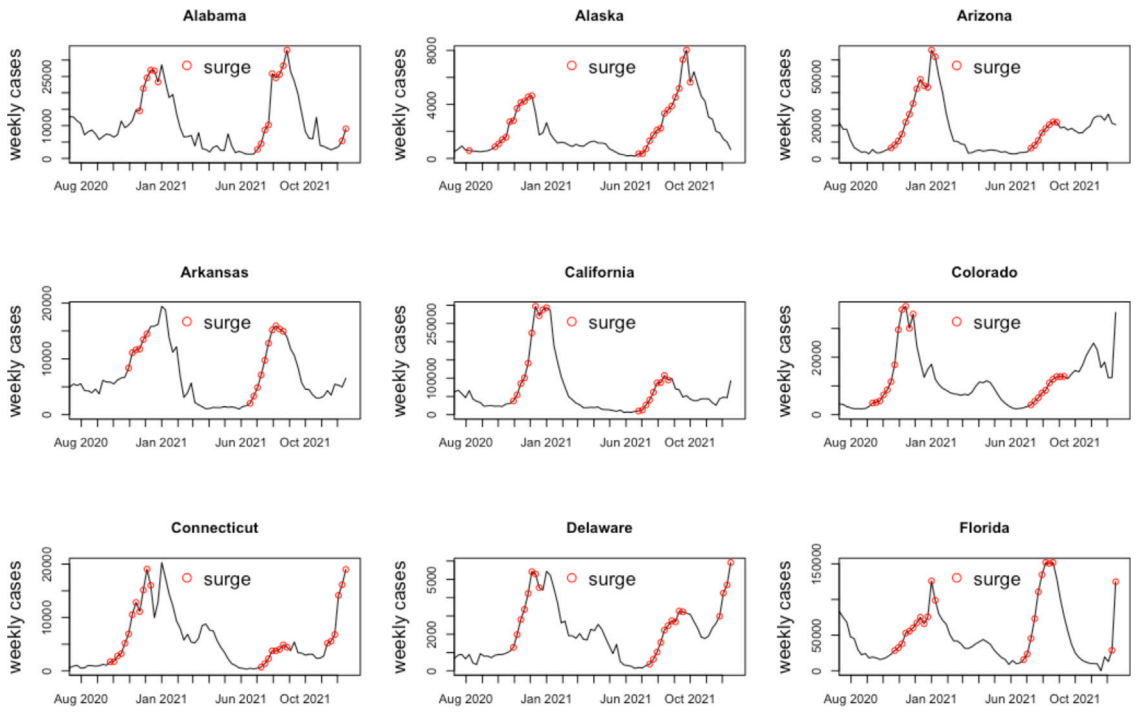

Fig. 2. Weekly reported cases in individual states with red points highlighting surge weeks, identified based on our criterion of >10% increase in smoothed weekly incidence sustained for more than three consecutive weeks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Results

4.1. Transmission density estimates

Estimates for transmission time parameters converged for all states in an average of 5 iterations using the alternating least-squares method described in Algorithm 1, applied to the first six months of Covid-19 case data for each state. Across the 50 states, the root mean squared error (RMSE) in predicted cases ranged from 6.0 cases/day in Vermont to 1,010 cases/day in Texas, with mean 173 cases. Louisiana had the highest population-adjusted RMSE of 9.0 cases/day per 100,000 persons.

As examples, the observed and modeled case counts for Texas and Louisiana are shown in Fig. 3 along with estimates from EpiNow2. Forecasts for the week after the training period are also shown, with both methods using the last estimated R_t to project cases beyond the observed period. High RMSE for fitted cases in these states is due to the model smoothing over several spikes in the observed case counts. Modeled cases from (1) and R_t estimates from (4) closely align with those from EpiNow2. EpiNow2 produces smoother R(t) estimates with lower uncertainty than the least-squares estimates due to its prior enforcing smoothness. Early in the pandemic, both methods show significant uncertainty and instability in $R(t)$. Standard errors are high when fewer than 1,200 cases had been reported, but decrease substantially as cumulative incidence increases. Eventually, R(t) fluctuates around 1, with estimates above 1 corresponding to periods of increasing incidence. Forecasts in incidence and R_t are also similar across models. On a MacBook Air with M1 chip, EpiNow2 required 6 min (Texas) and 3 min (Louisiana) to sample four chains, while least-squares estimates were computed in seconds.

The estimated truncated normal transmission densities for all 50 states are shown in Fig. 4. With the standard deviation fixed at 1, most states have a mean transmission time of around 6.2 days, while 10 states have means below 6. In Florida, Michigan, Missouri, and New York, the transmission means ranged from 1.1–1.5 days, possibly reflecting overfitting to random fluctuations in the case data.

The estimated background rate of immigration μ was less than 15 cases/day in 43 of the states. Louisiana had the highest background rate of 67 cases/day. The estimated parameters for each state are reported in supplementary Table 1.

4.2. Forecasting surges in cases

Fig. 5 shows ROC curves for forecasts of surges, averaged over all 50 states, using 3 methods: Hawkes estimates $\hat{R}_t > x$, an "incidence-based" model that uses the percent change in weekly incidence, and a random guesser that predicts surge weeks at random. The random guesser has a slope of two because a prediction is counted as correct if a surge occurs either one or two weeks after time t. The ROC curves for individual states using the Hawkes model are also shown in Fig. 5. Based on the averaged ROC curves, at a false positive rate of 1%, the Hawkes model achieves a true positive rate of 44%, compared to 10% for the incidence-based model and 2% for the random guesser. These rates correspond to a decision boundary of $R_t > 1.33$, and change in incidence $> 78\%$. With this false positive rate, the Hawkes model correctly predicts an average of 7 more surge weeks per state than the incidence-based model. The incidence-based model captures more true positives than the Hawkes model at false positive rates above 8%. The Hawkes model achieves its highest accuracy across all states of 90% using the decision boundary $R_t > 1.39$ while the highest average accuracy from incidence based model is 82% using percent change $> 35\%$.

Fig. 6 compares ROC curves using R_t estimates from Epinow2 in Georgia and Vermont, where the Hawkes model had the best and worst predictive performance, respectively. The ROC curves for both methods roughly align, with Hawkes estimates showing slightly better performance in Georgia, while Epinow2 performs better in Vermont. For each state, the 72 forecasts were calculated in less than a minute using the least-squares estimator (4). In contrast, obtaining posterior estimates for R(t) in a single state using EpiNow2 required 15 h. We sampled four chains on a supercomputer with eight cores, using a random walk prior over weekly intervals and the fixed transmission parameters estimated in Section 3.3.

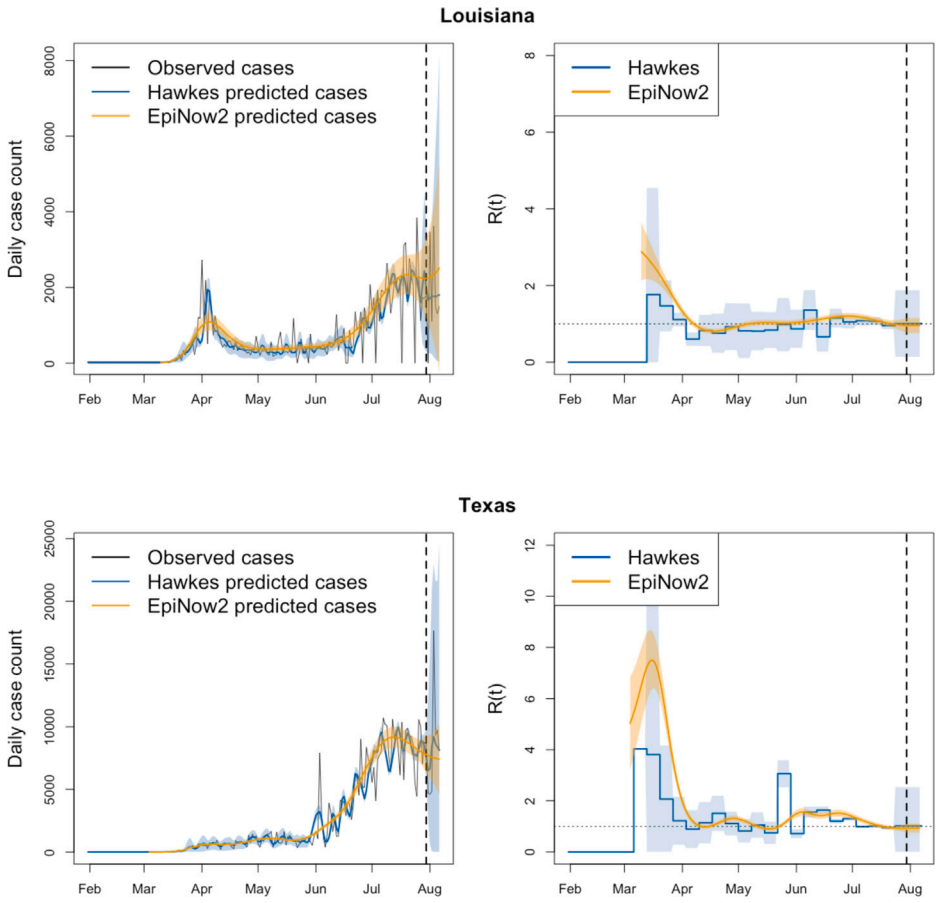

Fig. 3. Left panel: Observed daily reported case counts in 2020 (black), with modeled counts from the Hawkes model (blue) and EpiNow2 (orange). The ribbon indicates the range of estimated cases using the estimated reproduction rate ± one standard deviation, which is calculated based on (5) for the Hawkes model. Right panel: The reproduction rate R(t) estimated weekly based on (4) (blue) and daily by EpiNow2 (orange). Ribbons indicate the interval ± one standard deviation around the estimates. The lower end of the uncertainty interval is truncated at 0 to exclude negative estimates. Uncertainty estimates are not displayed for the Hawkes model until the cumulative case count exceeds 1,200. EpiNow2 does not calculate estimates until the first case. For reference, a dotted horizontal line indicates $R = 1$. In either panel, the vertical dashed line indicates the end of the observation period; values to the right of this line represent forecast estimates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

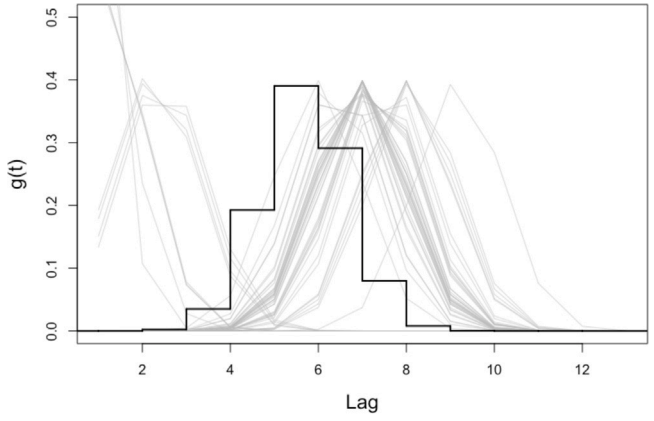

Fig. 4. Estimated transmission densities for all 50 states (gray lines), and averaged across all states (thick black line). The transmission density here reflects the time between reported cases and is assumed to follow a truncated normal distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Six states stand out with significantly lower true positive rates in Fig. 5. Three of these states, Florida, Michigan, and Missouri, had estimated transmission means around one, suggesting the model generalized poorly. The other three states, Vermont, Nevada, and Idaho, experienced periods of elevated or increasing incidence that did not meet our surge criteria, unlike most states where case counts quickly decreased after peaking. Fig. 7 illustrates this by showing the predicted surge weeks for Nevada and Vermont, which had the lowest true positive rates among all 50 states. Vermont experienced a prolonged period of sustained incidence in the summer of 2020 that was not correctly forecast by the Hawkes model, while Nevada saw a small increase in incidence in April 2021 that led to false positives in both models. For

Algorithm 1 Alternating least-squares procedure for simultaneous estimation of transmission parameters ν, baseline rate μ, and effective reproduction rate R_t as a step function over τ-length intervals.

Input: Incident count data $\{N_t\}_{t=1}^{T}$, transmission density function $g(t;\nu)$, exponential smoothing rate α, maximum iterations maxiter, convergence threshold ϵ.
Output: Estimated parameters $\hat{R}_0, \dots, \hat{R}_{T/\tau}$, $\hat{\nu}$, and $\hat{\mu}$
Initialize $R_0^{(0)}, \dots, R_{T/\tau}^{(0)}$, $\mu^{(0)}$, and $\nu^{(0)}$.
for each iteration $k = 1, 2, \dots,$ maxiter **do**
  Update $\nu^{(k)}$ by minimizing the root mean squared error:

$$\nu^{(k)} = \arg\min \sqrt{\mathrm{MSE}(\nu; R^{(k-1)}, \mu^{(k-1)})}$$

  where

$$\mathrm{MSE}(\nu) = \sum_{t=1}^{T} \left(N_t - \mu - \sum_{s=1}^{t-1} R_s N_s g(t-s;\nu) \right)^2.$$

  Form the $(T-1) \times (T/\tau - 1)$ matrix $G^{(k)}$, where:

$$G_{ij} = \sum_{s=(j-1)\tau+1}^{j\tau} g(i-s;\nu^{(k)}) N_s.$$

  Augment $G^{(k)}$ to include a column of ones to estimate an intercept:

$$G = (1 \quad G^{(k)}).$$

  Solve for $\mu^{(k)}$ and $R_0^{(k)}, \dots, R_{T/\tau-1}^{(k)}$ from G, incrementally increasing the regularization parameter ρ until:

$$\min \hat{R} > -0.01.$$

  Apply exponential smoothing to $R_0^{(k)}, \dots, R_{T/\tau-1}^{(k)}$ to estimate $R_{T/\tau}^{(k)}$
  if

$$\frac{|\nu^{(k)} - \nu^{(k-1)}|}{\nu^{(k-1)}} < \epsilon, \quad \frac{|\mu^{(k)} - \mu^{(k-1)}|}{\mu^{(k-1)}} < \epsilon, \quad \max\{R^{(k)} - R^{(k-1)}\} < \epsilon$$

  then
    break
  end if
end for
Return $R_0^{(k)}, \dots, R_{T/\tau}^{(k)}$, $\nu^{(k)}$, and $\mu^{(k)}$.

the other 44 states, the Hawkes model forecast very accurately: at a false positive rate of 5%, the average true positive rate of the Hawkes model among these states is 78%.

Fig. 8 shows the predicted R_t and change in incidence for each of the 72 forecast weeks across all 50 states. A linear SVM is used to determine the optimal separating line between the surge and non-surge weeks based on the previous week's change in incidence and predicted R_t. This optimal boundary achieves 82% accuracy at forecasting surge weeks, with a 4.9% false positive and 61% true positive rate.

5. Discussion

The Hawkes model provides a simple analytic formula for both real-time and retrospective estimation of the effective reproduction number R_t. Estimates can be calculated without distributional assumptions on R_t and accommodate a wide range of transmission densities. Our estimates align closely with those from EpiNow2 (Fig. 3), but are computed significantly more efficiently and do not require tuning priors or an MCMC sampler. At a false positive rate of 1%, the real-time Hawkes model had a 44% true positive rate over all 50 states, compared to a 10% true positive rate for an incidence-based model. This represents the correct detection of 7 more surges per state, on average, compared to forecasts based solely on observed incidence during the 72 forecast periods in this study (Fig. 5).

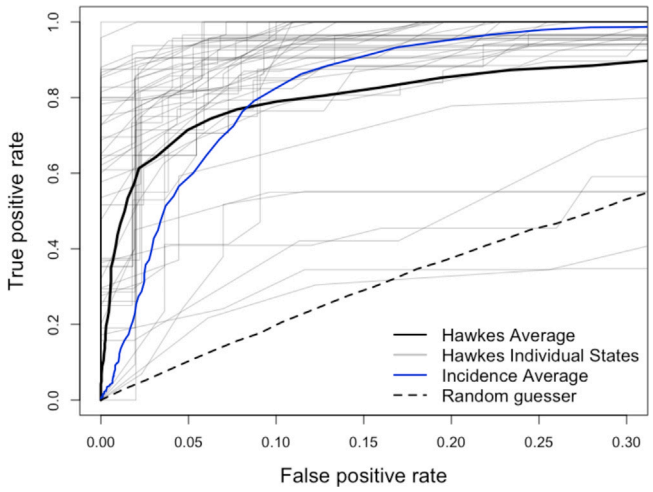

Fig. 5. ROC curves comparing three methods for forecasting surges in cases: R_t estimates from a Hawkes model (solid black), observed changes in incidence (blue), and a random predictor (dashed black). The true and false positive rates are averaged across all 50 states, with individual state curves shown for the Hawkes model (gray). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We estimate considerable variation in R_t over the study period (Fig. 3), reflecting both endemic seasonal patterns and changes in infectiousness due to factors such as the emergence of new variants and public health measures reducing the number of social contacts. Some methods attempt to disentangle these influences, such as the endemic-epidemic modeling framework (Held et al., 2005, 2006; Bracher and Held, 2022) which estimates a time-varying reproduction and baseline rate, often using sine-cosine terms to capture seasonality. For COVID-19 data, Congdon (2022) and Alaimo Di Loro et al. (2024) model a variable baseline and reproduction rate using space–time random effects while also incorporating spatial interactions. Future work should further develop methods for jointly estimating endemic and epidemic components of observed incidence, especially for emerging diseases like COVID-19 where historical data are limited. This could be achieved through parametric models incorporating covariates (Chiang et al., 2022) or non-parametric approaches such as Gaussian processes (Mohler, 2013).

The estimates of mean transmission means here (Fig. 4) are somewhat higher than in some other studies of Covid-19 transmission (Lauer et al., 2020; Schoenberg, 2023). The transmission time distribution estimated here reflects the intervals between reported cases, delayed from the actual infection times by the incubation period, which Lauer et al. (2020) estimated to have a median of 5.1 days, and potentially further lagged by delays in testing after exposure. Four states had transmission means around one day, potentially due to autocorrelation in daily case counts arising from reporting practices. Although, consistent with Schoenberg (2023), we did not find significant day-of-week effects in reported cases, the near exponential transmission distributions in these states indicate possible reporting issues. Future work should consider adjusting the baseline rate to account for reporting differences between weekdays and weekends that may bias transmission estimates. For simplicity, we assumed the transmission time distribution was fixed over the study period, but future work could extend this by using different transmission parameters in different time periods.

The least-squares uncertainty estimates (5) are unstable at the start of the pandemic when fewer than 1,200 cases were reported, but decrease significantly once more data are available. This instability is due to the sparsity in the initial columns of the transmission matrix whose inverse is used to calculate standard errors. Future work should consider methods to pool information across states to stabilize

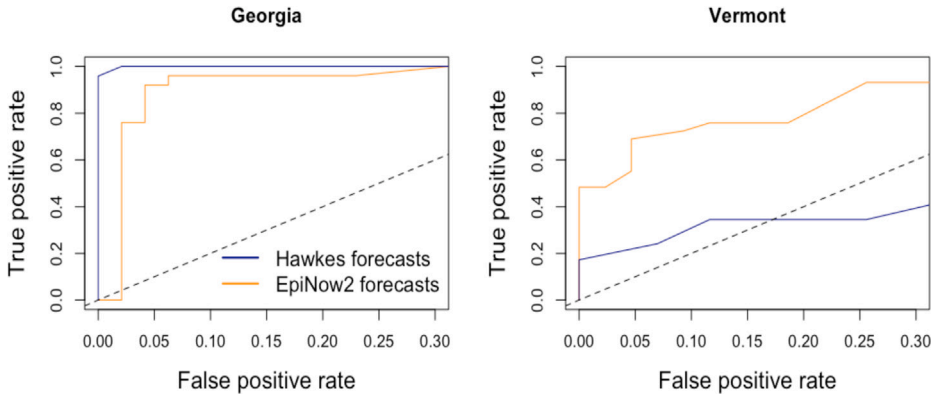

Fig. 6. ROC curves comparing the performance of R_t estimates from the Hawkes model (blue) and from EpiNow2 (orange) for predicting surges in cases in Georgia and Vermont. The dashed line indicates the performance of a random guesser. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

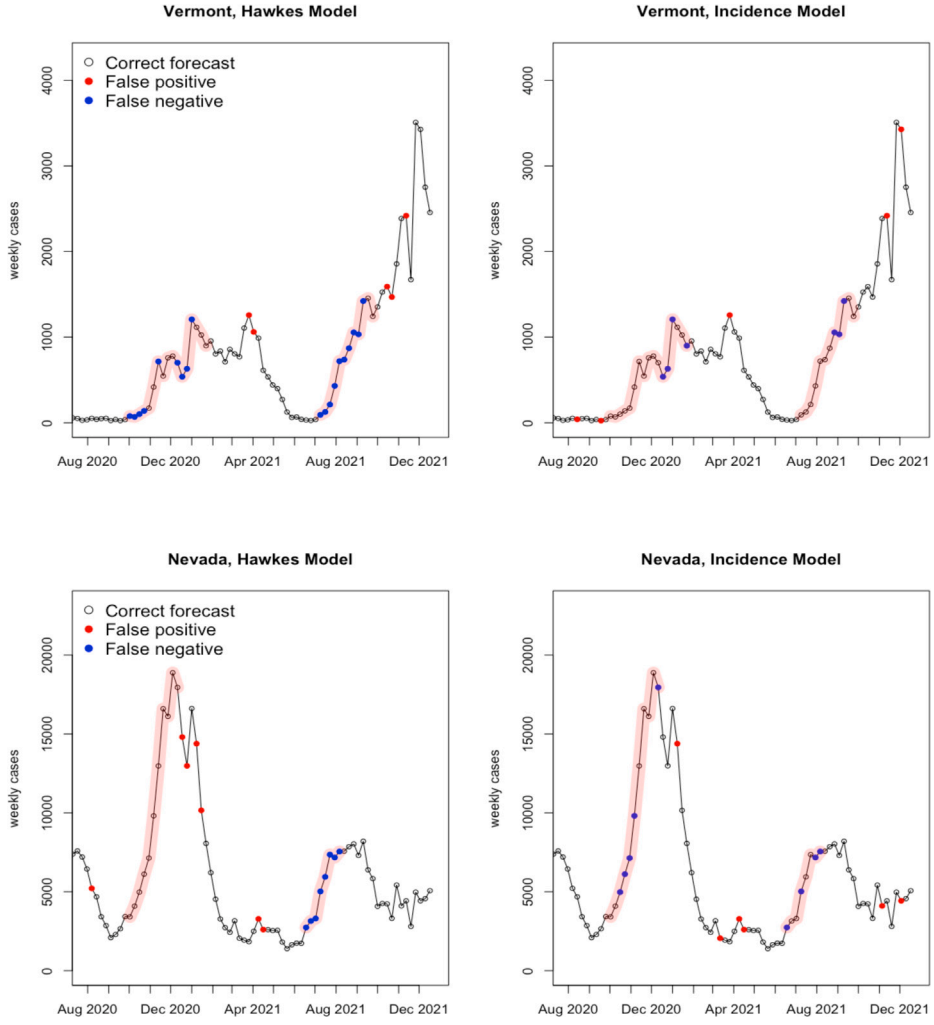

Fig. 7. Predicted surge weeks for the Hawkes model (left) and incidence-based model (right) at a false alarm rate of 10%. The corresponding decision boundaries are $R_t > 1.03$ and change in incidence > 26% for Vermont and $R_t > 1$ and change in incidence > 27% for Nevada. The red highlighted ribbon indicates true surge periods. Red dots indicate false alarms while blue dots indicate surge weeks that were missed by the model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

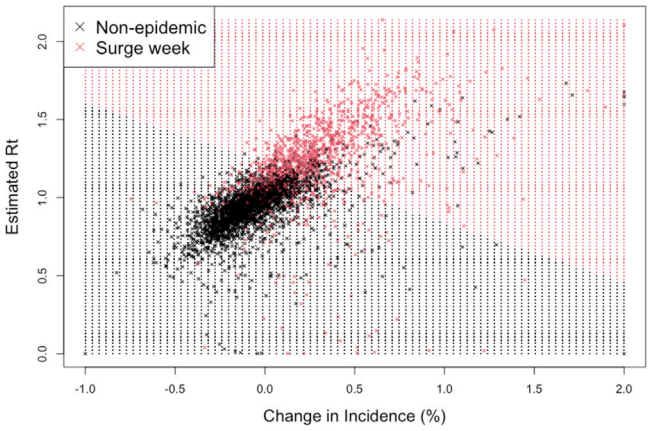

Fig. 8. Linear decision boundary between surge weeks (red) and non-surge weeks (black) determined by a linear SVM using data from all 50 states. Each marked point (x) indicates the percent change in incidence and the estimated R_t from the previous week, colored by whether the following week was during a surge. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

estimates, especially when incidence is low or towards the end of the observation window, without giving unreasonable weight to estimates from other states (Diouane et al., 2024). Additionally, methods to propagate uncertainty in the transmission parameters should be considered in future work.

CRediT authorship contribution statement

Sophie Phillips: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **George Mohler:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Frederic Schoenberg:** Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Acknowledgments

This material is based upon work supported by the National Science Foundation, United States under grant number DMS-2124433.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.epidem.2025.100824.

Data availability

The data used are publicly available from https://github.com/CSSEGISandData/COVID-19.

References

Abbott, S., Hellewell, J., Sherratt, K., Gostic, K., Hickson, J., Badr, H., DeWitt, M., Azam, J., Funk, S., 2024. EpiNow2: Estimate real-time case counts and time-varying epidemiological parameters. R package version 1.6.1, https://epiforecasts.io/EpiNow2/dev.

Alaimo Di Loro, D., Bohning, D., Sahu, S.K., 2024. A Bayesian spatio-temporal Poisson auto-regressive model for the disease infection rate: application to Covid-19 cases in England. J. R. Stat. Soc. Ser. C. Appl. Stat. qlae067. http://dx.doi.org/10.1093/jrsssc/qlae067.

Anderson, R., 1991. Infectious Diseases of Humans: Dynamics and Control. Cambridge University Press.

Bertozzi, A.L., Franco, E., Mohler, G., Short, M.B., Sledge, D., 2020. The challenges of modeling and forecasting the spread of Covid-19. Proc. Natl. Acad. Sci. 117 (29), 16732–16738. http://dx.doi.org/10.1073/pnas.2006520117.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory, Pittsburgh, 27-29, vol. 1992, pp. 144–152. http://dx.doi.org/10.1145/130385.130401.

Bracher, J., Held, L., 2022. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. Int. J. Forecast. 38 (3), 1221–1233. http://dx.doi.org/10.1016/j.ijforecast.2020.07.002.

Brett, T.S., Rohani, P., 2020. Dynamical footprints enable detection of disease emergence. PLoS Biol. 18, e3000697. http://dx.doi.org/10.1371/journal.pbio.3000697.

Brown, R.G., 1963. Smoothing, Forecasting and Prediction of Discrete Time Series. Prentice-Hall.

Browning, R., Sulem, D., Mengersen, K., Rivoirard, V., Rousseau, J., 2021. Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of Covid-19. PloS One 16 (4), e0250015. http://dx.doi.org/10.1371/journal.pone.0250015.

Cauchemez, S., Boelle, P.Y., Donnelly, C.A., Ferguson, N.M., Thomas, G., Leung, G.M., Hedley, A.J., Anderson, R.M., Valleron, A.J., 2006. Real-time estimates in early detection of SARS. Emerg. Infect. Dis. 12 (1), 110. http://dx.doi.org/10.3201/eid1201.050593.

Domenech de Cellès, M., Magpantay, F., King, A., Rohani, P., 2018. The impact of past vaccination coverage and immunity on pertussis resurgence. Sci. Transl. Med. 10 (434), http://dx.doi.org/10.1126/scitranslmed.aaj1748.

Centers for Disease Control (CDC), 2021. 2021 case definition. https://ndc.services.cdc.gov/case-definitions/coronavirus-disease-2019-2021/. (last Accessed 9 Dec 24).

Centers for Disease Control (CDC), 2024. Behind the model: CDC's tools to assess epidemic trends. https://www.cdc.gov/cfa-behind-the-model/php/data-research/rt-estimates/index.html. (last Accessed 28 Jan 2025).

Chen, Z., Dassios, A., Kuan, V., Lim, J.W., Qu, Y., Surya, B., Zhao, H., 2021. A two-phase dynamic contagion model for Covid-19. Results Phys. 26, 104264. http://dx.doi.org/10.1016/j.rinp.2021.104264.

Chiang, W.H., Liu, X., Mohler, G., 2022. Hawkes process modeling of Covid-19 with mobility leading indicators and spatial covariates. Int. J. Forecast. 38 (2), 505–520. http://dx.doi.org/10.1016/j.ijforecast.2021.07.001.

Congdon, P., 2022. A spatio-temporal autoregressive model for monitoring and predicting Covid infection rates. J. Geogr. Syst. 24 (4), 583–610. http://dx.doi.org/10.1007/s10109-021-00366-2.

Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S., 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. 178 (9), 1505–1512. http://dx.doi.org/10.1093/aje/kwt133.

Cowling, B.J., Wong, I., Lai-Ming, H., Riley, S., Leung, G.M., 2006. Methods for monitoring influenza surveillance data. Int. J. Epidemiol. 35 (5), 1314–1321. http://dx.doi.org/10.1093/ije/dyl162.

Diouane, Y., Schoenberg, F., Mohler, G., 2024. Accurate estimation of cross-excitation in multivariate Hawkes process models of infectious diseases. In: 11th IEEE International Conference on Data Science and Advanced Analytics. DSAA 2024.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track Covid-19 in real time. Lancet Infect. Dis. 20 (5), 533–534. http://dx.doi.org/10.1016/S1473-3099(20)30120-1.

Dong, E., Ratcliff, J., Goyea, T.D., Katz, A., Lau, R., Ng, T.K., Garcia, B., Bolt, E., Prata, S., Zhang, D., Murray, R.C., Blake, M.R., Du, H., Ganjikhanloo, F., Ahmadi, F., Williams, J., Choudhury, S., Gardner, L.M., 2022. The Johns Hopkins university center for systems science and engineering Covid-19 dashboard: data collection process, challenges faced, and lessons learned. Lancet Infect. Dis. 22 (12), e370–e376. http://dx.doi.org/10.1016/S1473-3099(22)00434-0.

Eichner, M., Dietz, K., 2003. Transmission potential of smallpox: Estimates based on detailed data from an outbreak. Am. J. Epidemiol. 158 (2), 110–117. http://dx.doi.org/10.1093/aje/kwg103.

Farrington, C.P., Kanaan, M.N., Gay, N.J., 2003. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 4 (2), 279–295. http://dx.doi.org/10.1093/biostatistics/4.2.279.

Ferguson, N.M., Cummings, D.A., Fraser, C., Cajka, J.C., Cooley, P.C., Burke, D.S., 2006. Strategies for mitigating an influenza pandemic. Nature 442 (7101), 448–452. http://dx.doi.org/10.1038/nature04795.

Harrigan, R., Mossoko, M., Okitolonda-Wemakoy, E., Schoenberg, F.P., Hoff, N., Mbala, P., Wannier, S.R., Lee, S.D., Ahuka-Mundeke, S., Smith, T.B., Selo, B., Njokolo, B., Rutherford, G., Rimoin, A.W., Tamfum, J.J.M., Park, J., 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. http://dx.doi.org/10.1016/j.epidem.2019.100354.

Hawkes, A., 1971. Spectra of some self-exciting and mutually exciting point processes. Biometrika 58 (1), 83–90. http://dx.doi.org/10.1093/biomet/58.1.83.

Held, L., Hofmann, M., Hohle, M., Schmid, V., 2006. A two-component model for counts of infectious diseases. Biostatistics 7 (3), 422–437. http://dx.doi.org/10.1093/biostatistics/kxj016.

Held, L., Hohle, M., Hofmann, M., 2005. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Stat. Model. 5 (3), 187–199. http://dx.doi.org/10.1191/1471082X05st098oa.

Johns Hopkins Center for Systems Science and Engineering (JHU CSSE), https://github.com/CSSEGISandData. (last Accessed 9 Sep 24).

Kaur, T., Sarkar, S., Chowdhury, S., Sinha, S.K., Jolly, M.K., Dutta, P.S., 2020. Anticipating the novel Coronavirus disease (COVID-19) pandemic. Front. Public Heal. 8 (521), http://dx.doi.org/10.3389/fpubh.2020.569669.

Kirchner, M., 2016. Hawkes and INAR(∞) processes. Stochastic Process. Appl. 26 (8), 2494–2525. http://dx.doi.org/10.1016/j.spa.2016.02.008.

Koyama, S., Horie, T., Shinomoto, S., 2021. Estimating the time-varying reproduction number of Covid-19 with a state-space method. PLoS Comput. Biol. 17 (1), 1–18. http://dx.doi.org/10.1371/journal.pcbi.1008679.

Kresin, C., Schoenberg, F.P., Mohler, G., 2022. Comparison of Hawkes and SEIR models for the spread of Covid-19. Adv. Appl. Stat. 74, 83–106. http://dx.doi.org/10.17654/0972361722019.

Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., 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. http://dx.doi.org/10.7326/M20-0504.

Liu, R., Zhong, J., Hong, R., Chen, E., Aihara, K., Chen, P., Chen, L., 2021. Predicting local COVID-19 outbreaks and infectious disease epidemics based on landscape network entropy. Sci. Bull. http://dx.doi.org/10.1016/J.SCIB.2021.03.022.

Ma, Q., Liu, J., Liu, Q., Kang, L., Liu, R., Jing, W., Wu, Y., Liu, M., 2021. Global percentage of asymptomatic SARS-CoV-2 infections among the tested population and individuals with confirmed Covid-19 diagnosis: A systematic review and meta-analysis. JAMA Netw. Open 4 (12), e2137257. http://dx.doi.org/10.1001/jamanetworkopen.2021.37257.

Meyer, S., Elias, J., Hohle, M., 2012. A space–time conditional intensity model for invasive meningococcal disease occurrence. Biometrics 68 (2), 607–616. http://dx.doi.org/10.1111/j.1541-0420.2011.01684.

Mohler, G., 2013. Modeling and estimation of multi-source clustering in crime and security data. Ann. Appl. Stat. 7 (3), 1525–1539. http://dx.doi.org/10.1214/13-AOAS647.

Mohler, G.O., Schoenberg, F.P., Short, M.B., Sledge, D., 2021. Analyzing the impacts of public policy on Covid-19 transmission: A case study of the role of model and dataset selection using data from Indiana. Stat. Public Policy 8 (1), 1–8. http://dx.doi.org/10.1080/2330443X.

O'Brien, D.A., Clements, C.F., 2021. Early warning signals predict emergence of COVID-19 waves. Biol. Lett. 17, http://dx.doi.org/10.1101/2021.06.24.

Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. J. Amer. Statist. Assoc. 83 (401), 9–27. http://dx.doi.org/10.2307/2288914.

Ogata, Y., 1998. Space-time point-process models for earthquake occurrences. Ann. Inst. Statist. Math. 50, 379–402. http://dx.doi.org/10.1023/A:1003403601725.

Park, J., Chaffee, A.W., Harrigan, R.J., Schoenberg, F.P., 2022. A non-parametric Hawkes model of the spread of Ebola in west Africa. J. Appl. Stat. 49, 621–637. http://dx.doi.org/10.1080/02664763.2020.1825646.

Phillips, S., Schoenberg, F., 2024. Efficient nonparametric estimation of variable productivity Hawkes processes. J. Appl. Stat. 1–18. http://dx.doi.org/10.1080/02664763.2024.2426019.

Rizoiu, S., Mishra, S., Kong, Q., Carman, M., 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. http://dx.doi.org/10.1145/3178876.3186108.

Schoenberg, F., 2022. Nonparametric estimation of variable productivity Hawkes processes. Environmetrics 33 (6), http://dx.doi.org/10.1002/env.2747.

Schoenberg, F., 2023. Estimating Covid-19 transmission time using Hawkes point processes. Ann. Appl. Stat. 17 (4), 3349–3362. http://dx.doi.org/10.1214/23-AOAS1765.

Schoenberg, F.P., Hoffmann, M., Harrigan, R.J., 2019. A recursive point process model for infectious diseases. Ann. Inst. Statist. Math. 71 (5), 1271–1287. http://dx.doi.org/10.1007/s10463-018-0690-9.

Son, W.S., Park, J., Kwon, O., 2020. Early detection of influenza outbreak using time derivative of incidence. EPJ Data Sci. 9 (1), 28. http://dx.doi.org/10.1140/epjds/s13688-020-00246-7.

Southall, E., Tildesley, M.J., Dyson, L., 2020. Prospects for detecting early warning signals in discrete event sequence data: Application to epidemiological incidence data. PLoS Comput. Biol. 16 (9), http://dx.doi.org/10.1371/journal.pcbi.1007836.

Unkel, S., Farrington, C.P., Garthwaite, P.H., Robertson, C., Andrews, N., 2012. Statistical methods for the prospective detection of infectious disease outbreaks: a review. J. R. Stat. Soc. 175, 49–82. http://dx.doi.org/10.1111/rssa.2011.175.issue-1.

Wallinga, J., Teunis, P., 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am. J. Epidemiol. 160, 509–516. http://dx.doi.org/10.1093/aje/kwh255.