UCLA
UCLA Electronic Theses and Dissertations

Title

Fitting Multivariate Hawkes Models to COVID-19 Data from All 50 States in the United States

Permalink

https://escholarship.org/uc/item/41p157rn

Author

Gong, Wanling

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
Los Angeles

Fitting Multivariate Hawkes Models to COVID-19 Data from All 50 States in the United States

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of in Applied Statistics And Data Science

by

Wanling Gong

2024

© Copyright by
Wanling Gong
2024

ABSTRACT OF THE DISSERTATION

Fitting Multivariate Hawkes Models
to COVID-19 Data from All 50 States
in the United States

by

Wanling Gong
Master of in Applied Statistics And Data Science
University of California, Los Angeles, 2024
Professor Frederic R. Paik Schoenberg, Chair

This paper investigates whether the distribution of SARS-CoV-2 (COVID-19) transmission times can be reliably estimated using only case count data, employing the Hawkes model as the analytical framework. Hawkes point processes, widely recognized for modeling and analyzing time-to-event data, offer a robust approach to understanding transmission dynamics. This study fits the Hawkes model with varying productivity levels to case count data from all 50 U.S. states. Transmission time density is estimated using nonparametric methods and normal approximations.

The findings indicate that, for most states, the mean transmission time is approximately 7 days, with a standard deviation of about 1 day (Science, 2020). These estimates are compared across states and with prior reports, revealing slightly shorter average transmission durations and reduced variability in this study. Furthermore, the results highlight that the virus can be transmitted as early as the first day of contact, emphasizing its potential for rapid spread (World Health Organization, 2020). As derived from this analysis, a deeper

understanding of SARS-CoV-2 transmission dynamics carries significant implications for public health modeling and policy-making (Pan et al., 2020).

The thesis of Wanling Gong is approved.

Maria Cha

Nicolas Christou

YingNian Wu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2024

To my parents, Qianming Gong and Weili Wan,

who among so many other things,

nurtured my curiosity and guided me with unwavering love.

To my husband Sidi Sun,

who supports me at all times, with endless patience and encouragement.

To everyone who was a part of this journey,

thank you for your kindness, inspiration, and support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 virus has profoundly impacted global society, reshaping daily life, healthcare systems, and economies. The World Health Organization (World Health Organization, 2020) noted that the pandemic has resulted in a massive loss of life and unprecedented challenges to public health, food systems, and employment. Economically, it has triggered the worst global recession since World War II, pushing millions into extreme poverty (World Bank, 2020). Healthcare systems are under tremendous strain, with service disruptions and increased workloads for health workers [World Health Organization, 2020]. Figure 1.1 depicts the trend in weekly COVID-19 deaths and total COVID-19 mortality per 100,000 population (age-adjusted) in the United States reported to the CDC. In that situation, Understanding the transmission dynamics of the virus, including the serial interval between consecutive cases and the incubation period from exposure to symptom onset, is critical to controlling the outbreak and developing effective strategies to reduce transmission.

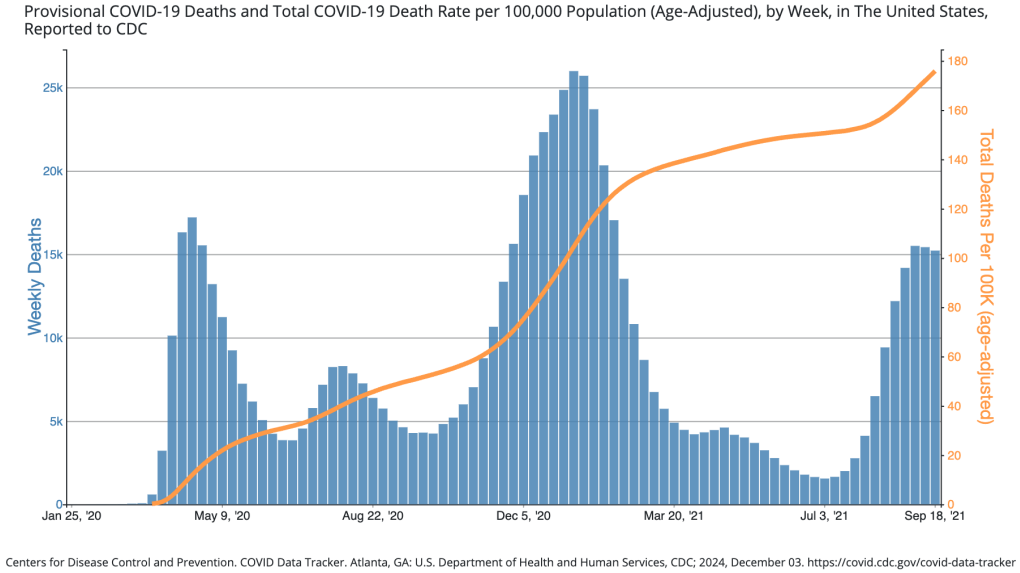

Figure 1.1: The Trend In Weekly COVID-19 Deaths and Total COVID-19 Mortality Per 100,000 Population

The emergence of the SARS-CoV-2 virus and the resulting COVID-19 pandemic have presented unprecedented challenges to global health systems, necessitating the development of accurate models to predict and manage the spread of the virus. Central to controlling the pandemic has been the ability to model its spread accurately. Epidemiological models, which aim to predict infection trends and inform public health interventions, rely on a comprehensive understanding of key transmission parameters. One such critical parameter is the transmission time—the interval between an individual contracting the virus and infecting others. This metric is fundamental for optimizing public health policies, including quarantine durations, contact tracing strategies, and resource allocation, yet remains underexplored in many contexts (Hethcote, 2000; Fraser et al., 2004).

Transmission time governs the dynamics of an epidemic's spread and determines the reproduction number key measure of how quickly a virus spreads through a population (Anderson & May, 1991). While the incubation period, defined as the time from exposure to symptom onset, has been extensively studied, the transmission interval adds a layer of

complexity.

Studies have determined that the incubation period for SARS-CoV-2 has a median duration of four to five days, extending up to 14 days in some cases (Guan et al., 2020; Lauer et al., 2020; Li et al., 2020). Approximately 97.5 % of symptomatic individuals show symptoms within 11.5 days of infection (Lauer et al., 2020). Research from Wuhan, China, revealed that the median time from symptom onset to the development of acute respiratory distress syndrome (ARDS) ranges between eight and 12 days, while the median time to ICU admission falls between 9.5 and 12 days (Huang et al., 2020; Wang et al., 2020; Yang et al., 2020; Zhou et al., 2020).

The Centers for Disease Control and Prevention (CDC) summarized these findings, noting that the incubation period spans two to 14 days (CDC, 2021a). For individuals with SARS-CoV-2, the contagious period typically lasts up to 10 days following symptom onset (CDC, 2021d) or 14 days after exposure (CDC, 2021c, 2021e), but it can extend to 20 days in some cases (CDC, 2021a). Accordingly, the CDC recommends a 14-day home isolation period after the last contact with an infected individual (CDC, 2021b, 2021c). Similarly, the World Health Organization (WHO) advises a 14-day quarantine period after exposure, citing an average incubation period of five to six days, with a possible range of up to 14 days (WHO, 2021a, 2021b). It reflects not only biological factors, such as viral load and infectiousness but also behavioral patterns and environmental influences, including population mobility and adherence to preventive measures (Ferretti et al., 2020). Understanding this metric is vital for projecting infection waves and designing targeted interventions such as vaccination campaigns or NPIs (non-pharmaceutical interventions) like mask mandates and physical distancing. Without precise estimates of transmission time, epidemiological models risk oversimplifying the nuanced dynamics of disease spread.

The COVID-19 pandemic, however, has demonstrated that the dynamics of disease transmission are far from uniform. Factors such as population density, healthcare access, and cultural norms significantly shape how the virus spreads within different regions. For in-

stance, densely populated urban centers, with frequent close-contact interactions and high mobility, have shown faster transmission rates compared to sparsely populated rural areas, where individuals naturally engage in fewer close-contact interactions (Nuzzo et al., 2020). Furthermore, differences in state-level policies, such as the timing of mask mandates, the implementation of school closures, or the availability of testing and vaccination, have created a patchwork of transmission patterns across the U.S. (Chernozhukov et al., 2021). These regional variations underscore the need for localized studies that account for such heterogeneity, particularly in large, diverse countries like the United States.

Hawkes processes, originally developed in seismology to model earthquake aftershocks (Hawkes, 1971; Ogata, 1988), offer a promising solution to these challenges. These models are uniquely designed to account for self-exciting processes, where the occurrence of one event increases the likelihood of subsequent events in a defined temporal and spatial window. Their adaptability has led to successful applications in diverse fields. In finance, they model high-frequency trading patterns, capturing rapid cascades of trades triggered by initial market shocks (Bowsher, 2007). In criminology, they have been used to analyze crime patterns, identifying hotspots where one incident increases the probability of additional crimes nearby (Mohler et al., 2011). Social media platforms have leveraged Hawkes processes to track the virality of posts, modeling how initial shares lead to exponential growth in reposts (Zhao et al., 2015). These sectoral applications illustrate the model's versatility in capturing cascading, event-driven dynamics.

In epidemiology, Hawkes processes offer distinct advantages over traditional models. Unlike SEIR frameworks, which rely on fixed compartments and static parameters, Hawkes models dynamically adjust to reflect real-time data and event dependencies (Schoenberg, 2013). This flexibility is particularly valuable for capturing superspreading events and temporal shifts in transmission dynamics, such as those driven by policy interventions or behavioral changes. For example, during the early months of the pandemic, the serial interval—the time between successive cases in a transmission chain—shortened in regions with stringent

NPIs, such as lockdowns or mask mandates (Ali et al., 2020; Sun et al., 2021). Hawkes processes excel at modeling these dynamic shifts, providing insights into how interventions shape the trajectory of an epidemic.

Despite their strengths, Hawkes processes are not without limitations. Their reliance on high-quality, real-time data can be a challenge in the context of COVID-19, where reporting delays, underreporting, and data inconsistencies are common (Zhuang et al., 2004). Moreover, extending these models to multivariate settings—where interactions between regions or populations are considered—poses computational and interpretive challenges, particularly when data dimensions grow large (Bacry et al., 2015). These limitations highlight the importance of thoughtful implementation and underscore the need for further methodological advancements to fully realize their potential in epidemic modeling.

This study builds on the growing body of research that applies Hawkes processes to epidemic dynamics by extending the analysis to a broader range of U.S. states. By incorporating state-level heterogeneity in parameters and capturing regional variations in transmission times, this research aims to provide deeper insights into how localized factors influence the spread of the virus. Such an approach not only improves our understanding of COVID-19 transmission dynamics but also contributes to the development of adaptable modeling tools that can inform responses to future pandemics.

CHAPTER 2

Data

In this study, it includes 50 states in the United States, and daily SARS-CoV-2 case surveillance data are obtained through the CDC's public website https://covid.cdc.gov/covid-data-tracker, CDC (2021f)). The data set used in this article contains the number of daily cases from January 23, 2020, to August 25, 2021, a total of 582 days. For good analysis, parameters were estimated in 36 non-overlapping windows, each containing 16 days. This approach allows 576 days of data to be used in the analysis, that is, the period from January 23, 2020, to August 19, 2021. Each window provides a detailed snapshot of SARS-CoV-2 case progression, enabling systematic analysis of temporal trends and changes. These data were made public by the CDC and downloaded on August 26, 2021. This comprehensive dataset provides a solid foundation for understanding the temporal dynamics of SARS-CoV-2 spread in the United States over a specified time frame.

Figure 2.1 illustrates the total number of reported cases across U.S. states during the 582-day observation period. The distribution of cases strongly correlates with state population sizes, as larger states such as California, Texas, and Florida report the highest case counts, reflecting their status as the most populous states in the nation. States with smaller populations, such as Wyoming and Vermont, naturally show much lower total case counts.

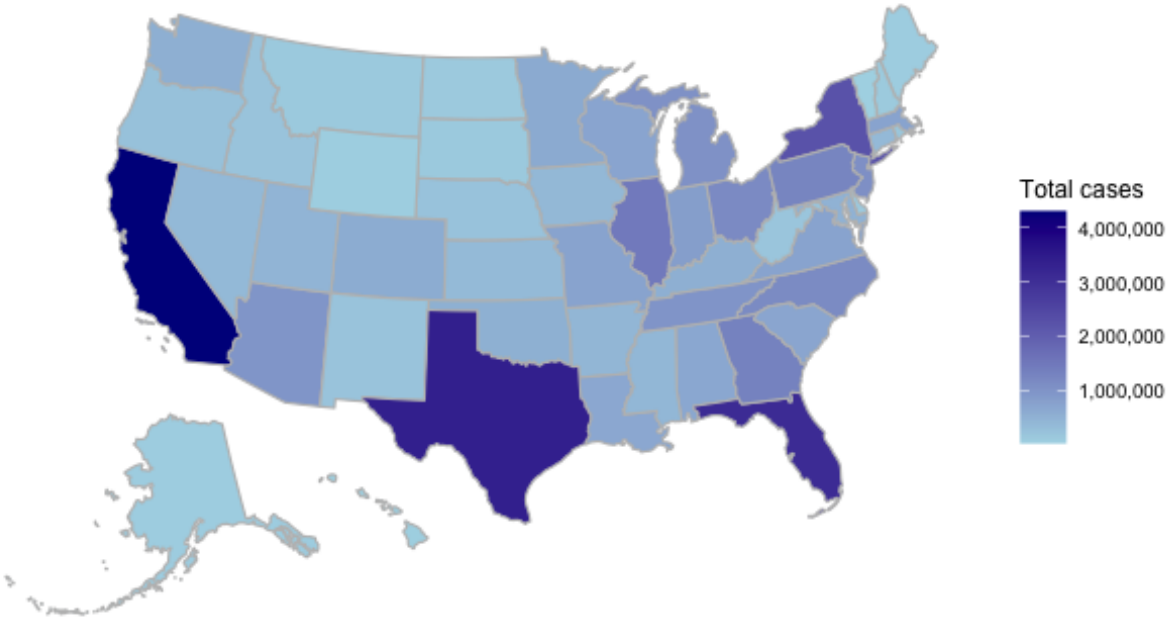

Figure 2.1: Total Number Of Reported Cases Across U.S. States

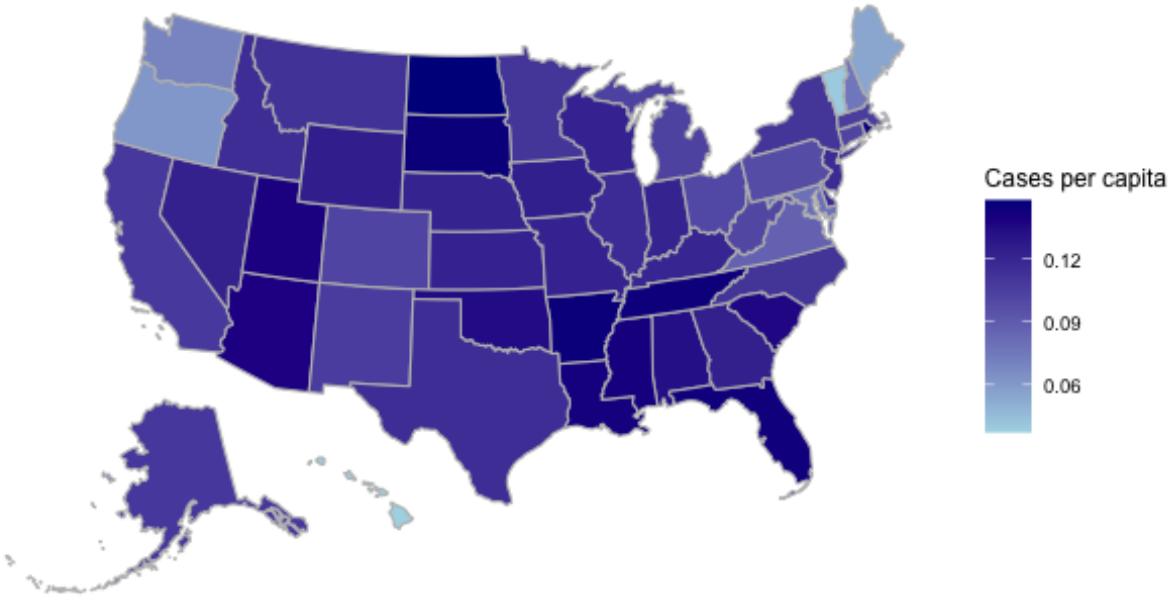

Figure 2.2: Per Capita Perspective By Displaying The Number Of Cases

Figure 2.2 provides a per capita perspective by displaying the number of cases relative to population size, using publicly available population data sourced from https://www.census.gov/data /tables/time-series/demo/popest/2020s-state-total.html. This perspective reveals important nuances. For instance, while California reports the highest absolute case numbers, its per capita incidence is moderate compared to smaller states like North Dakota (0.1489), Rhode Island (0.1471), South Dakota (0.1468), and Arkansas (0.1465), which exhibit the highest cases per capita. This suggests that smaller states with fewer total cases can still experience higher per capita burdens, likely influenced by factors such as population density, public health responses, and demographic structures.

The data also show that most states exhibit relatively consistent per capita rates of recorded incidence, with 41 out of 50 states falling within the range of 0.10 to 0.15. This clustering reflects a common baseline of disease spread across much of the U.S. However, four states—Hawaii (0.0380), Vermont (0.0401), Maine (0.0548), and Oregon (0.0625)—stand out for their exceptionally low per capita incidence[4]. These states may have benefited from geographic isolation, lower population density, or effective public health measures that mitigated the spread of the disease. The contrast between total case counts and per capita rates highlights the importance of considering population-adjusted metrics when assessing disease burden across regions.

The Centers for Disease Control and Prevention (CDC) aggregates COVID-19 data submitted voluntarily by state and territorial health departments. This reporting process is standardized but not mandatory, as the CDC does not possess direct authority to compel data submissions. Instead, it relies on cooperative agreements with jurisdictions. Data discrepancies can arise due to differences in reporting protocols, definitions of cases, and the timing of updates, leading to variations between CDC-reported figures and those presented by state or local health department websites (CDC FAQ).

The COVID-19 totals compiled by the CDC include both confirmed and probable SARS-CoV-2 cases and deaths, following the criteria outlined in the Council of State and Territorial

Epidemiologists (CSTE) case definitions. However, some exceptions apply, such as individuals repatriated from Wuhan, China, and Japan CDC (2021f), which were excluded from state and jurisdictional totals (CDC Technical Notes)

The reported numbers for a given day reflect the information provided by states and jurisdictions, which may correspond to the actual date of case occurrence or the date the case was reported. Such differences in methodologies contribute to temporal and geographic variability in case data (CSTE Position Statements).

This passage highlights key limitations and considerations in analyzing SARS-CoV-2 case data. The dates associated with recorded cases often differ significantly from the actual onset of the disease. In this analysis, the "transmission time" refers to the interval between the recorded dates of two cases. This interval reflects not only the time for disease incubation and expression but also the variability in reporting times. If cases resulting from rapid transmissions are more likely to be recorded than those with delayed transmissions, the average transmission time may be underestimated.

Missing data is a significant challenge in SARS-CoV-2 studies, as estimating the number of unreported cases remains highly complex (Bertozzi et al., 2020; Kresin, Schoenberg, and Mohler, 2021). During the early stages of the pandemic, the CDC conducted extensive seroprevalence studies in the spring and summer of 2020 to estimate the virus's prevalence in various locations through random sampling and testing (Bajema et al., 2021). However, these rigorous studies were discontinued following funding cuts to the CDC by the Trump administration in the summer of 2020 (Wermer and Stein, 2020). While missing data rates may vary by state (Bajema et al., 2021), no specific states are identified as having particularly unreliable data.

For further details about the CDC's SARS-CoV-2 case surveillance data collection, readers are referred to CDC Surveillance FAQ or CDC (2021).

Figure 2.3 visualizes reported case trends per capita across six representative states, re-

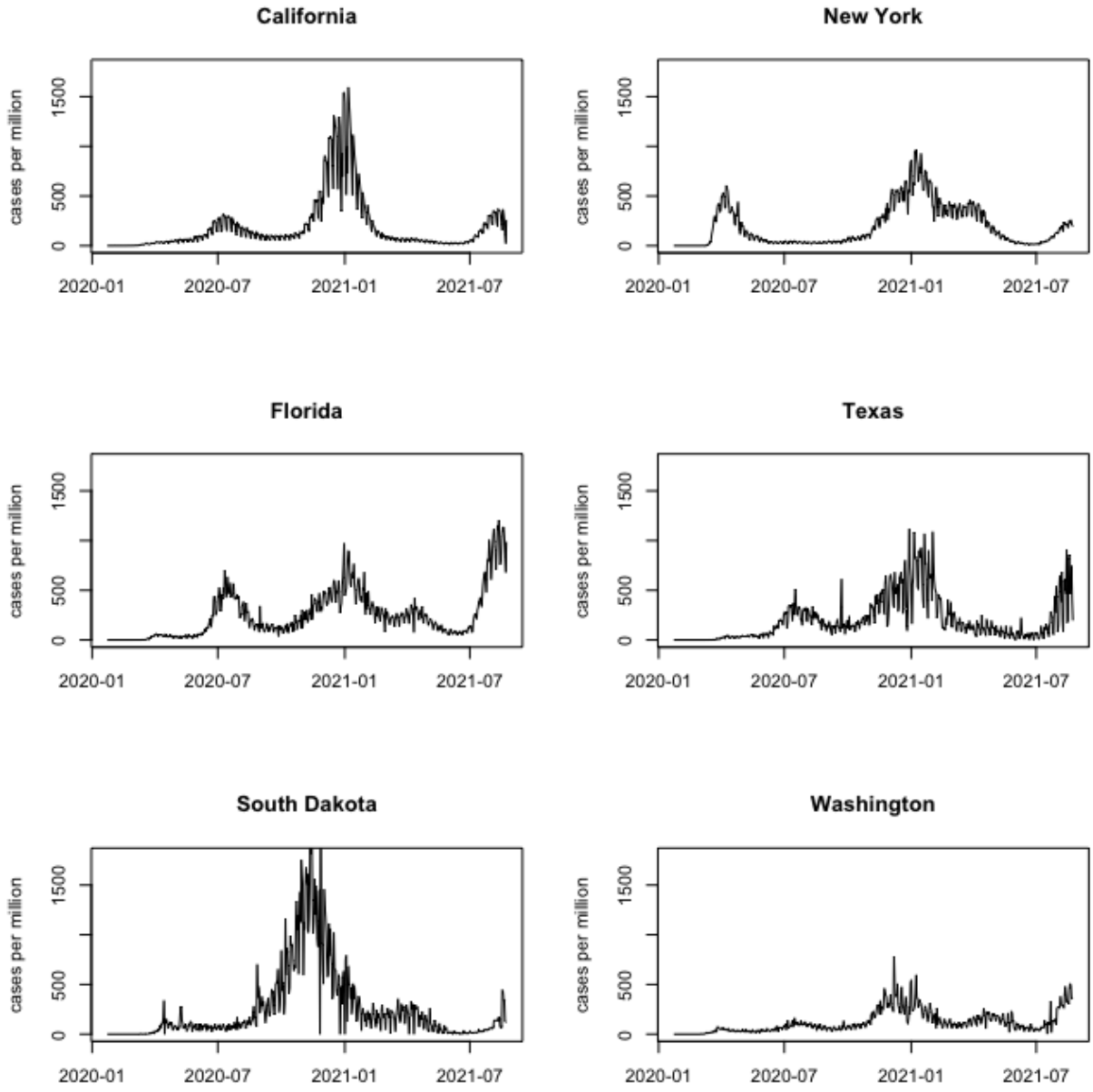

Figure 2.3: Case Trends Per Capita Across Six Representative States

vealing the sharp, nearly exponential patterns of increase and decline typical of epidemic models like SEIR and Hawkes models (Rizoiu et al., 2018; Kresin, Schoenberg, and Mohler, 2021). These trends reflect the dynamic nature of the pandemic and the influence of transmission dynamics and reporting practices.

CHAPTER 3

Methods

3.1 Overview

Hawkes point process models are self-exciting stochastic processes that model time-dependent events, where one event increases the probability of subsequent events occurring nearby. They are well suited for analyzing SARS-CoV-2 (COVID-19) dynamics because they capture the temporal and spatial clustering consistent with the spread of infection, where each case increases the probability of other cases that are close in time or space. We explain what Hawkes point process models are and how to use them in the following sections.

The Hawkes point process model has been utilized to estimate the temporal distribution of the spread of SARS-CoV-2 (COVID-19) based on the number of daily cases reported in all 50 states of the United States. This approach defines the number of cases as a self-exciting point process, where new infections arise from a combination of background events (independent spontaneous infections) and triggering events (infections caused by previous cases). By modeling these temporal dynamics, the study identifies connections and properties of disease transmission.

3.2 Methods

3.2.1 Hawkes Point Process Model

The Hawkes or self-exciting point process model (Hawkes, 1971) is a statistical framework used to model clustered point patterns, such as those observed in seismology, finance, crime, and infectious disease spread (Cauchemez, 2006, Daley, 2003, Ogata, 1988, Reinhart, 2018). The model describes the conditional intensity function $\lambda(t, x, y)$, which represents the expected rate at which events (e.g., confirmed cases) occur over time t, given the history of prior events:

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

where:

- μ: A constant *background rate* representing external, non-triggered events.
- $K(t)$: Represents *productivity*, the expected number of secondary events triggered directly by a single event at time t.
- $g(\cdot)$: A *density function* (non-negative and integrates to 1), referred to as the *triggering density* or *transmission time density*, describing the time-dependent effect of an event on subsequent occurrences.

The productivity $K(t)$ is a key parameter that captures the ability of events to trigger additional events and is closely tied to the reproduction number (R_t) in epidemic models such as SEIR (Bertozzi (2020), Kresin (2021)). For a Hawkes process with $0 < K < 1$, the expected number of events triggered by an individual event is:

$$K + K^2 + K^3 + \cdots = \frac{K}{1 - K}.$$

This relationship also implies that the fraction of events that are background (non-triggered) is $1-K$. The conditional intensity $\lambda(t, x, y)$ fully characterizes the finite-dimensional distribution of the point process (Prop. 7.2.IV of Daley and Vere-Jones, 2003).

3.2.2 Parameter Estimation

3.2.2.1 Precise Event Time Data

When the precise times of individual events are available, parameters of the Hawkes model are typically estimated using *Maximum Likelihood Estimation (MLE)*. This method provides statistical guarantees, including asymptotic unbiasedness, consistency, and efficiency (Ogata, 1978). The background rate μ and the triggering function g can be modeled parametrically or non-parametrically. Non-parametric approaches allow flexibility in capturing variations in g and μ, as shown in studies by ((Marsan and Lengline, 2008), (Zhuang, 2004)). Bayesian methods have also been explored for parameter estimation and uncertainty quantification ((Mohler,2013), (Rasmussen, 2013)).

3.2.2.2 Daily Aggregated Data

When only aggregated daily event counts are available, precise event times cannot be used. In such cases, we estimate parameters using a *least squares approach*, minimizing the sum of squared differences between observed and expected daily counts. This approach leverages the connection between Hawkes processes and *autoregressive time series models* ((Kirchner, 2016), (Kirchner, 2017)), allowing for efficient computation while preserving the interpretability of the model.

3.2.3 Model Implementation

Simulations demonstrate that the estimation of the transmission time distribution using Equation (3) is highly accurate. The simulated processes follow the methodology detailed in Section 3.3 of Reinhart (2018) or outlined in Section 1 of the Supplementary Material (Schoenberg, 2023). In each simulation, background points are initially generated via a homogeneous Poisson process with a rate μ. Each background point triggers a random number of additional points, determined by a Poisson random variable with mean K. These triggered points are distributed in time according to the density g relative to their triggering point. The process continues recursively, with each point potentially triggering further points until no new points are generated within the 576-day observation window.

Figure 3.1 illustrates the actual and estimated transmission time densities for simulated Hawkes processes. These models employ three distinct normal densities with varying means and standard deviations. Across the 576-day simulations, the estimated transmission time distributions closely match the true distributions. The estimated transmission time distributions have higher peaks while the tails of true distributions are longer. Parameters for the first set of simulations include $\mu = 1$ point/day, $K = 0.95$, $\nu = 9$ days, and $\sigma = 1$ day. For two additional simulations, μ and K remain unchanged, while (ν, σ) are set to (11 days, 2 days) and (4 days, 1.2 days), respectively.

For the scenario with $\mu = 1$ point/day, $K = 0.95$, $\nu = 9$, and $\sigma = 1$, 50 simulations were conducted, each spanning 576 days. Compared to Figure 3.2, Figure 3.3 provides a more precise estimate for stats. The resulting actual and estimated transmission time densities are displayed in Figure 3.3. Root mean square (RMS) errors for ν and σ across the 50 simulations were 0.148 and 0.224, respectively, demonstrating the accuracy of the least squares estimates for scenarios akin to SARS-CoV-2 transmission data.

Figure 3.4 depicts a modified scenario where 10% of cases occurring on Saturdays and 20% of cases occurring on Sundays are recorded on the following Monday. Despite this

recording bias, the mean estimated ν was 8.92 (true $\nu = 9.0$), and the mean estimated σ was 1.56 (true $\sigma = 1.0$). However, the RMS errors for ν and σ increased to 0.198 and 0.576, respectively, indicating a moderate impact of recording errors on the least squares estimates.

3.2.3.1 Parametric Model

In the parametric version of the model, the triggering density $g(u)$ is assumed to follow a normal distribution:

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

where ν and σ^2 represent the mean and variance of the triggering density, respectively. The parameter vector θ for the parametric model includes μ, ν, σ^2, and $K(t)$, with $K(t)$ estimated for each time interval. This results in a total of 39 parameters.

3.2.3.2 Non-Parametric Model

In the non-parametric model, the triggering density $g(u)$ is approximated by a step function defined over 16-day intervals. The step heights are constrained to sum to 1, ensuring $g(u)$ remains a valid probability density function. This reduces the number of free parameters associated with g to 15, resulting in a total of 52 parameters for the non-parametric model.

For both models, parameters are estimated by minimizing the following objective function:

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

where:

- $N(t)$: The observed number of events (e.g., confirmed cases) on the day t,

- $T = 576$: The total number of days in the dataset (36 time windows of 16 days each).

3.2.4 Optimization Procedure

Parameter optimization is performed using the *Nelder-Mead algorithm* implemented in R's `optim` function. Initial values for μ, ν, σ^2, and $K(t)$ are derived from prior estimates, and the maximum number of iterations is set to 100,000 to ensure convergence. For the non-parametric model, step heights for $g(u)$ are initialized iteratively, using the ending values from one iteration as the starting values for the next.

3.2.5 State-Specific Analysis

To account for heterogeneity across geographical regions, we fit the Hawkes model independently for each state, allowing all parameters to vary across states. This approach enables the model to capture region-specific dynamics in the underlying processes while maintaining consistency in methodology across states.

3.2.6 Summary

By fitting both parametric and non-parametric versions of the Hawkes model, we aim to provide a comprehensive analysis of the dynamics of event occurrences. The flexibility of the model ensures robust parameter estimation across different data granularities and geographical regions, while the optimization process ensures computational efficiency and reliability.

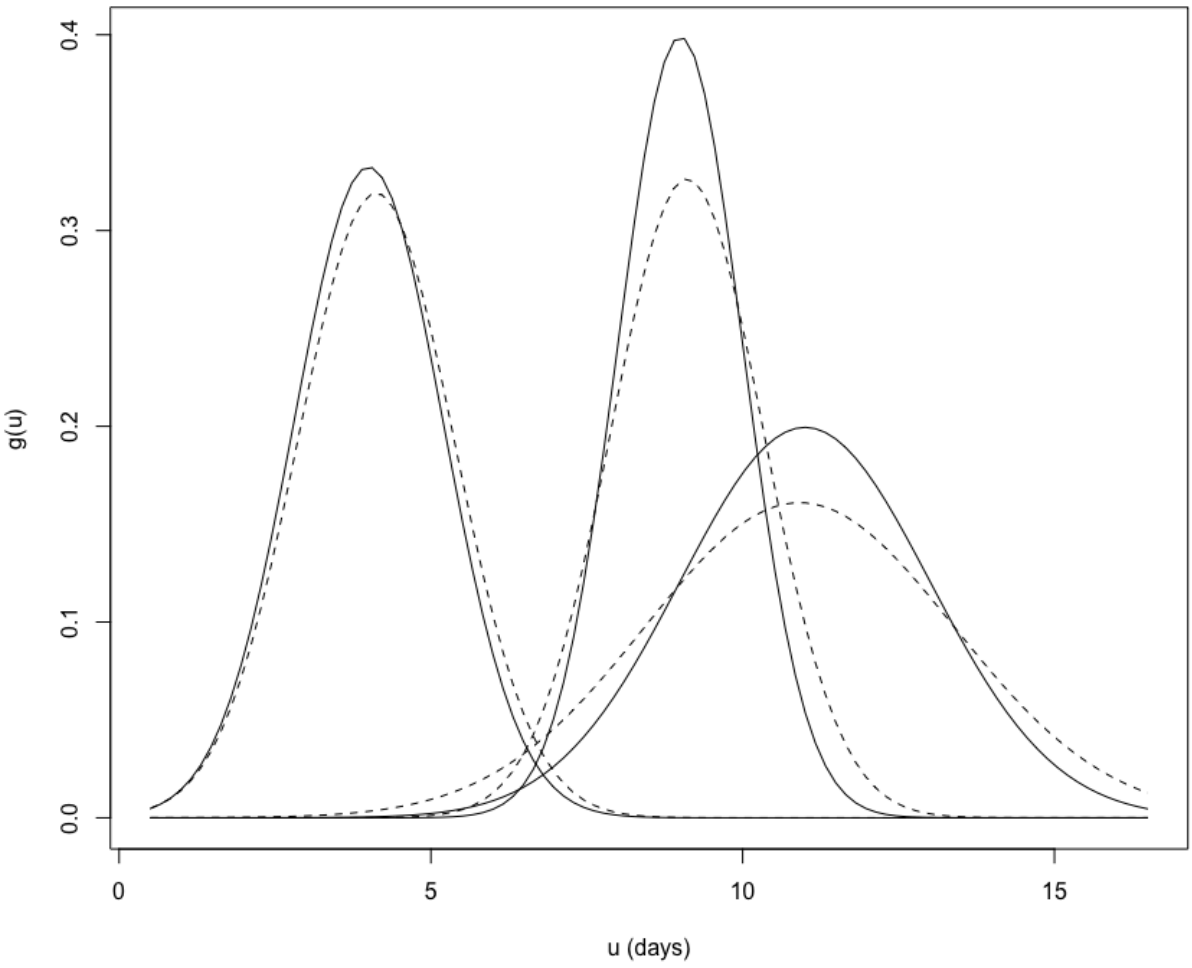

Figure 3.1: The Real and Estimated Transmission Time Densities For Simulated Hawkes Models

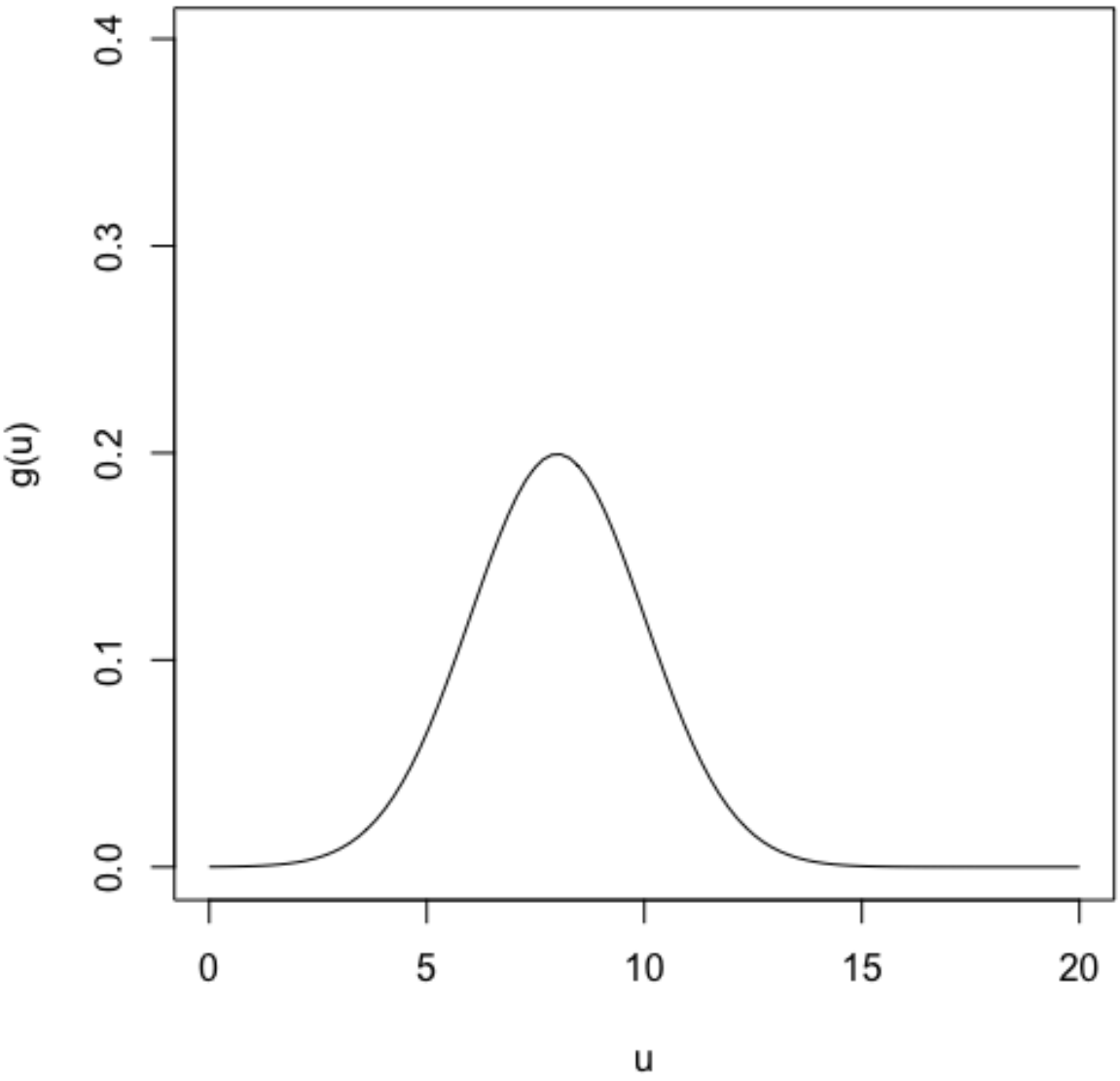

Figure 3.2: Time Densities for Arizona

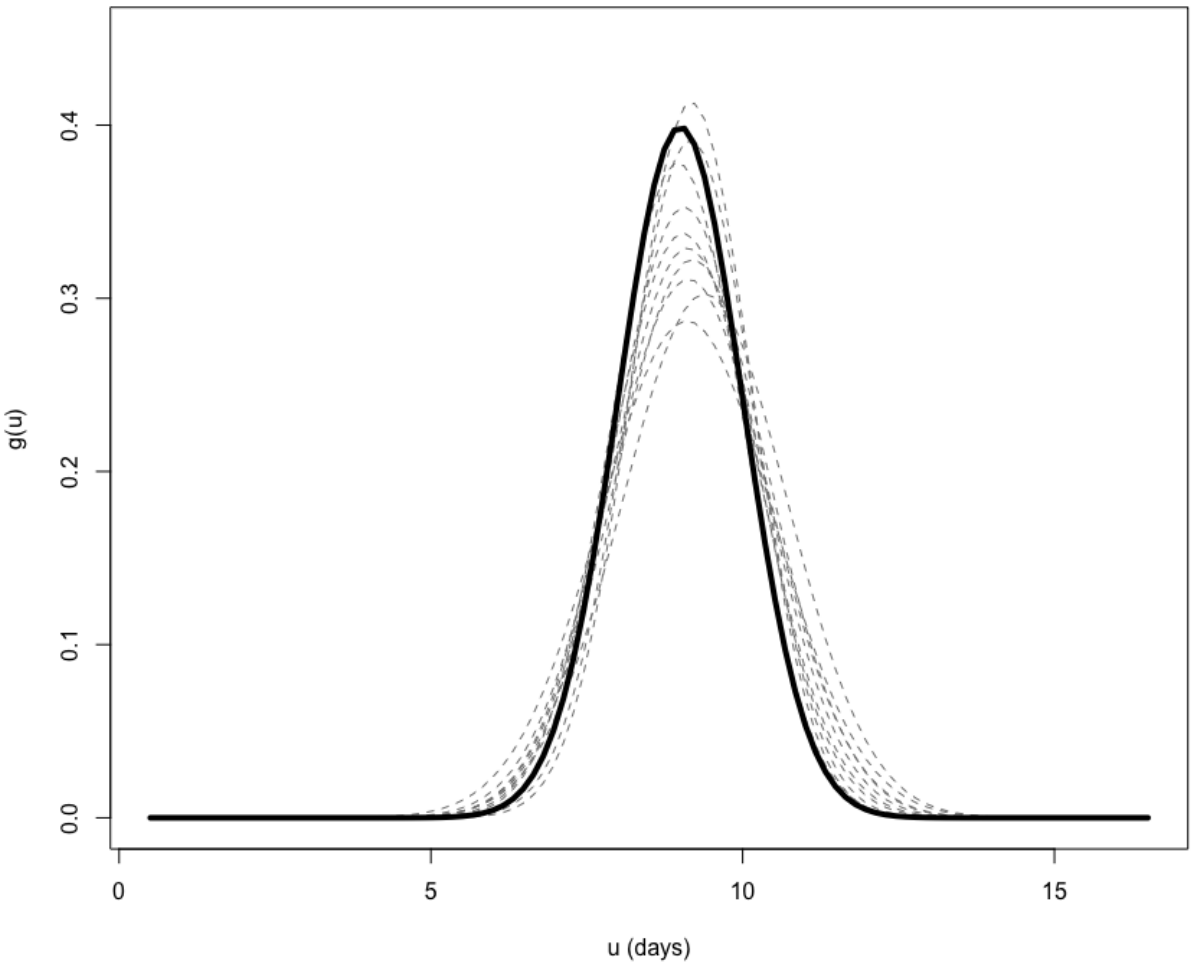

Figure 3.3: The Real And Estimated Transmission Time Densities

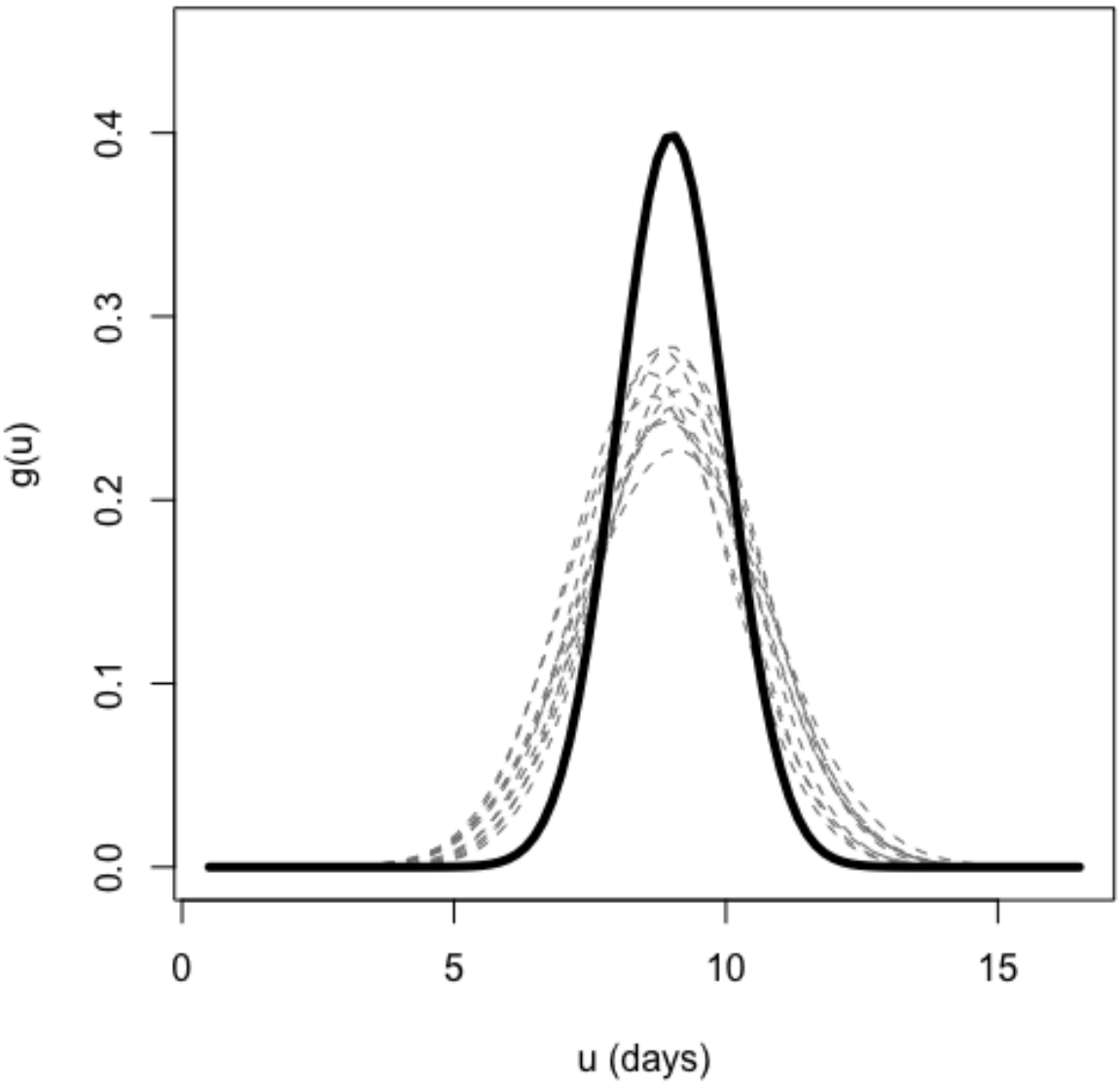

Figure 3.4: The Real And Estimated Transmission Time Densities With Event Dates

CHAPTER 4

Result

The Hawkes model provides an excellent fit for case-count data across all 50 states. For instance, Figure 4.1 illustrates how closely the Hawkes model, using a nonparametric estimate of the transmission density, approximates observed case counts in California, New York, Nevada, and Maryland. It also displays the corresponding estimated productivities, $K(t)$, for the model fitted to each of these states. The root mean square (RMS) errors in daily case counts for California, New York, Nevada, and Maryland are 2699, 861, 226, and 341, respectively. Across all 50 states, the RMS errors ranged from 22.1 in Vermont to 2729 in Texas, with a median error of 355.3 in New Jersey, a mean error of 493.7, and a standard deviation of 539.4. States with larger populations and higher case counts generally exhibited larger RMS errors. (Schoenberg, 2024)

Figures 4.1 - 4.4 present the nonparametric transmission time density estimates for all 50 states and their mean. Figure 4.1 shows the close approximation of the Hawkes model with non-parametrically estimated transmission density to the truly observed case counts throughout the observation period for California, New York, Nevada, and Maryland as well as the corresponding estimated productivity $K(t)$ for the Hawkes model fit data from each of these four states. Although the approximation is very close to reality, Hawkes's model fails to capture the severe fluctuation of case count especially for Maryland.

Figures 4.2 and 4.3 show a clear peak at seven days, suggesting a weekly cycle in the transmission dynamics. Also, there are masses at 1 day and 14 days, which indicates within 16 16-day periods, 1 day, 7 days, and 14 days are the strong triggering times. While the normal density offers a reasonable approximation of the transmission time distribution, the nonparametric estimates also show some mass at one-day and 14-day transmission times. However, as shown in Figure 4.4, no strong weekly cycle is apparent in the confirmed case counts by weekday, either across states or overall. Kansas is an exception, displaying significant variability in case counts by weekday, with elevated counts on Wednesdays and Fridays and lower counts on Tuesdays and Thursdays—likely due to reporting practices.

Using the normal approximation for transmission time, Figure 4.5 illustrates the estimated transmission time densities for each state alongside the mean. There are several outliers of which the normal distribution is shifted to the left a lot suggesting the unusually low estimated mean transmission time, though the reason for this remains unclear.

Figures 4.6 and 4.7 provide estimates of ν and σ for the fitted normal transmission time densities in all 50 states. Estimates of ν exhibit strong consistency across most states, with values ranging from 6.51 to 7.22 days for 46 states, and standard errors between 0.10 and 0.21 days. However, four outliers—Ohio, Virginia, Oklahoma, and Kansas—deviate notably. The estimated σ values are also consistent across states, ranging between 0.406 and 1.56 days, with standard errors from 0.15 to 0.27 days. Kansas is an obvious outlier, which may stem from reporting irregularities, such as numerous days with zero confirmed cases.

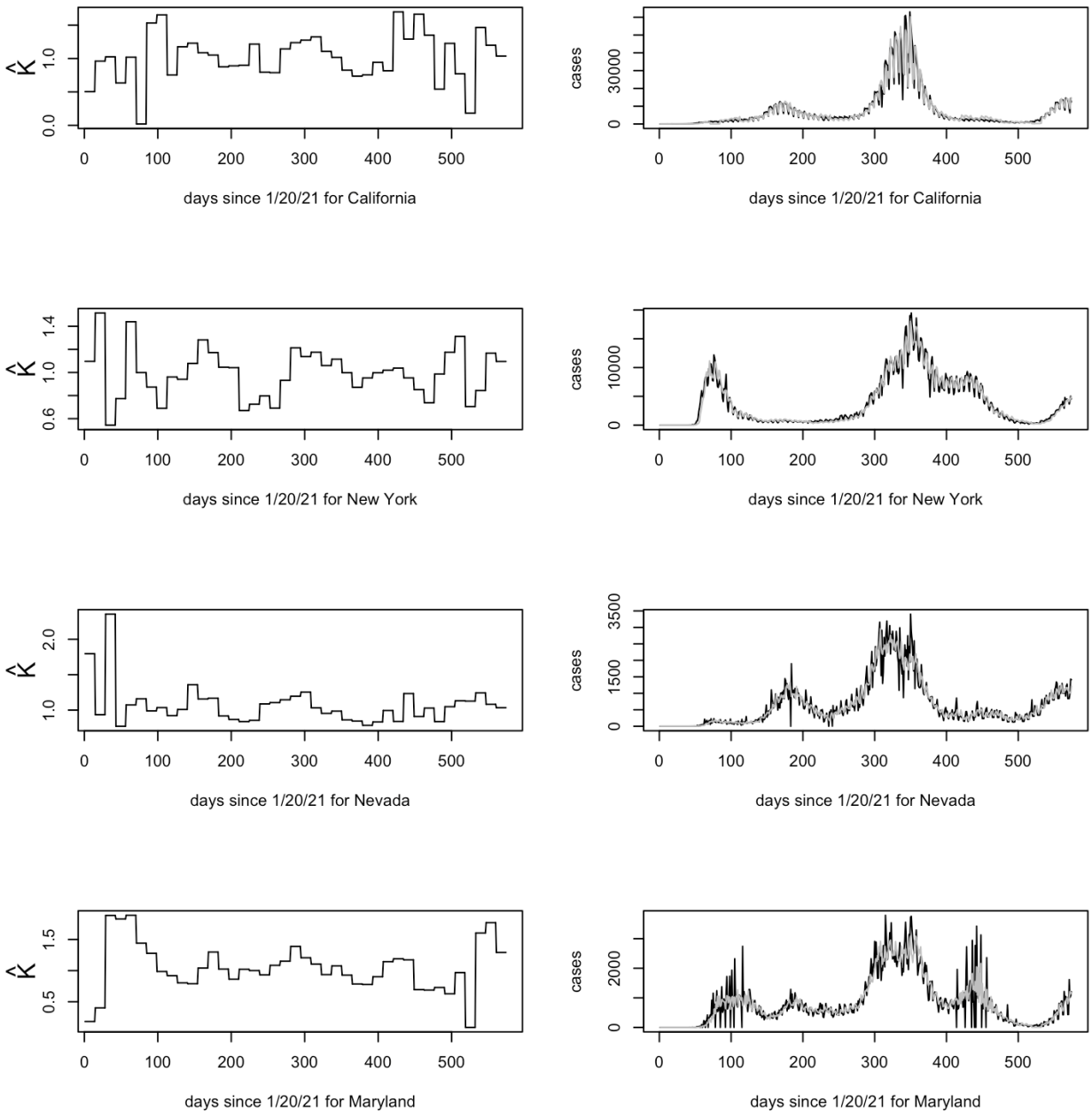

Figure 4.1: Nonparametric Transmission Time Density Estimates For 4 States

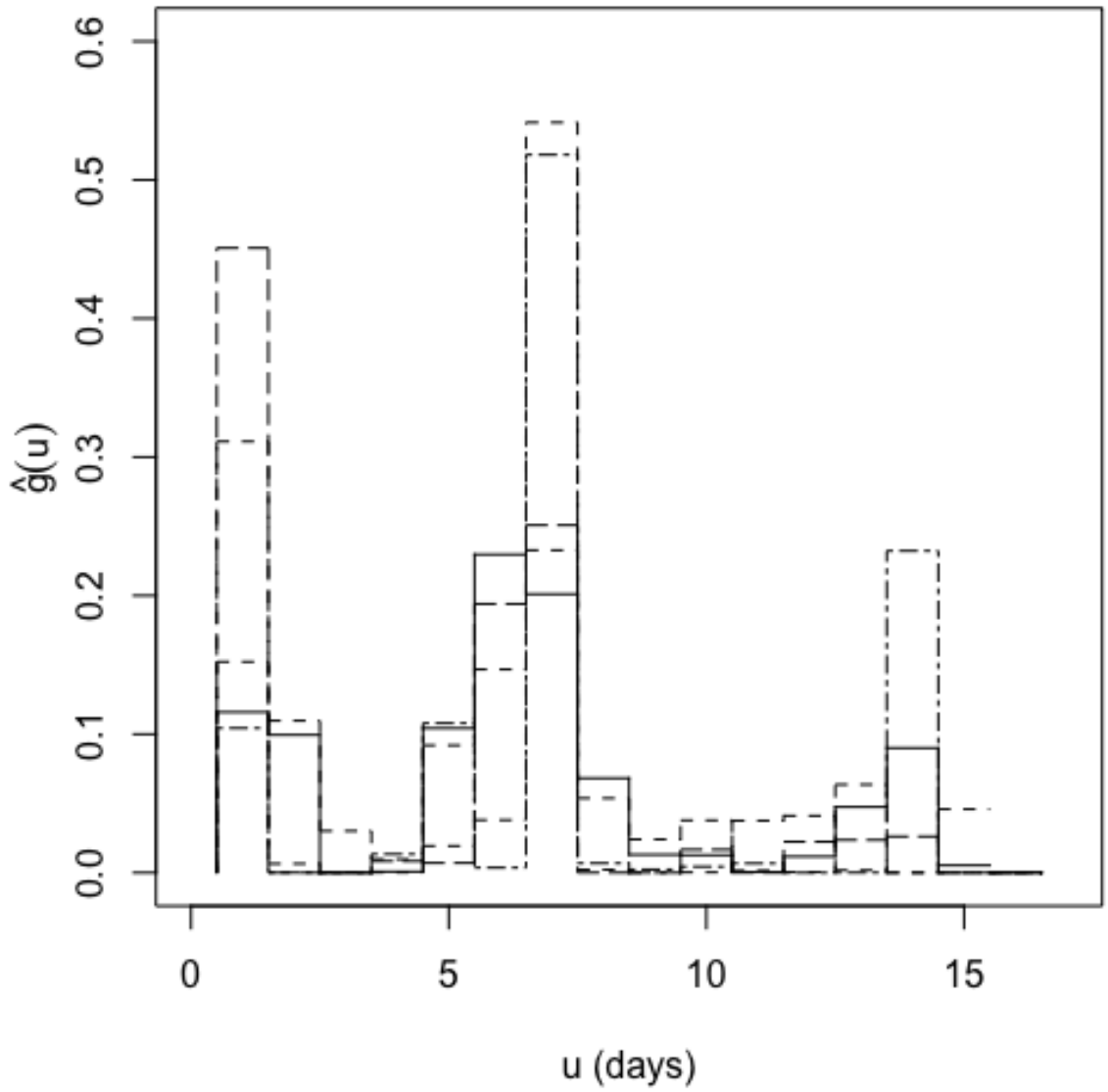

Figure 4.2: Nonparametric Transmission Time Density Estimates For 9 States

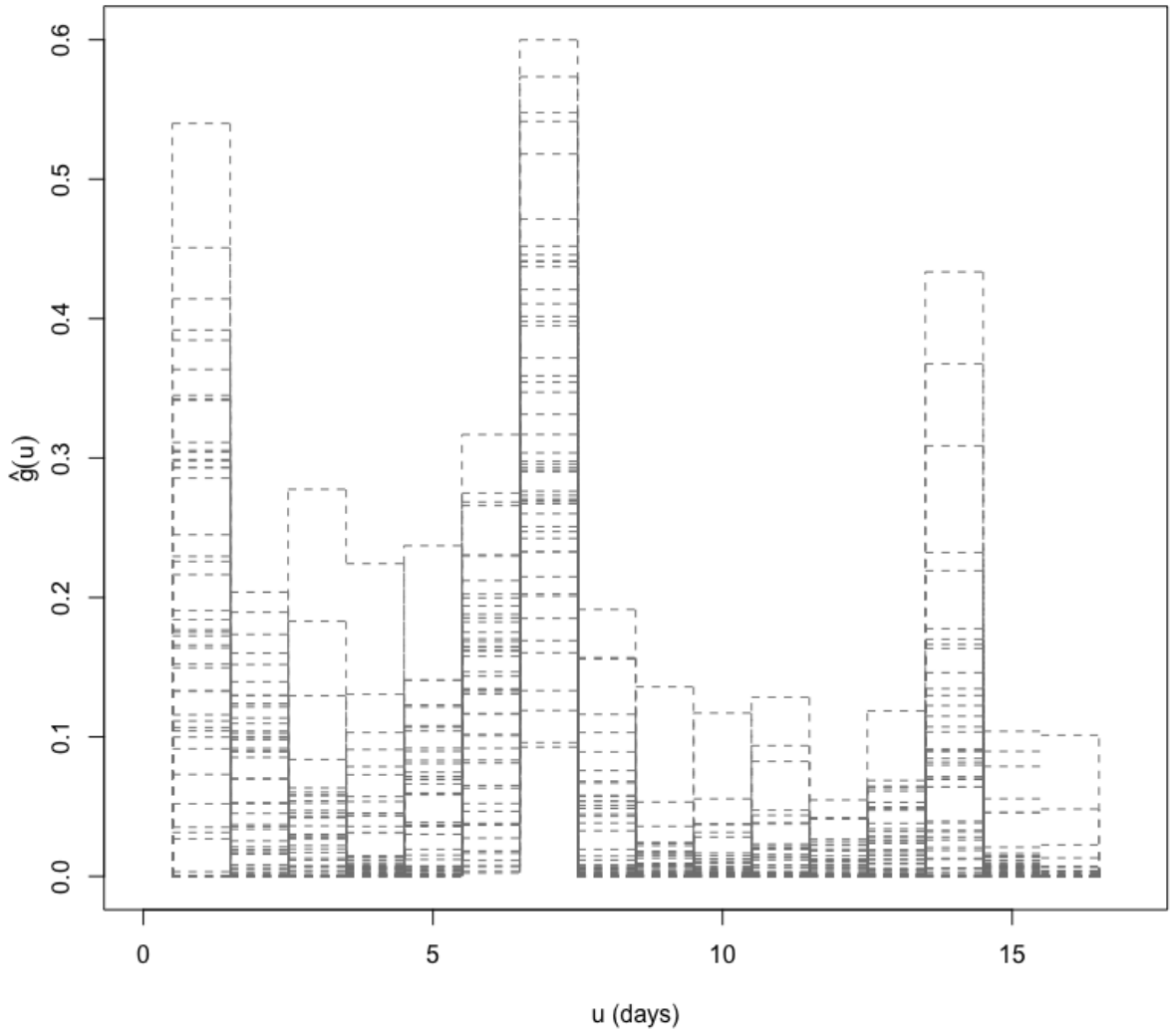

Figure 4.3: Nonparametric Transmission Time Density Estimates And Mean For All 50 States

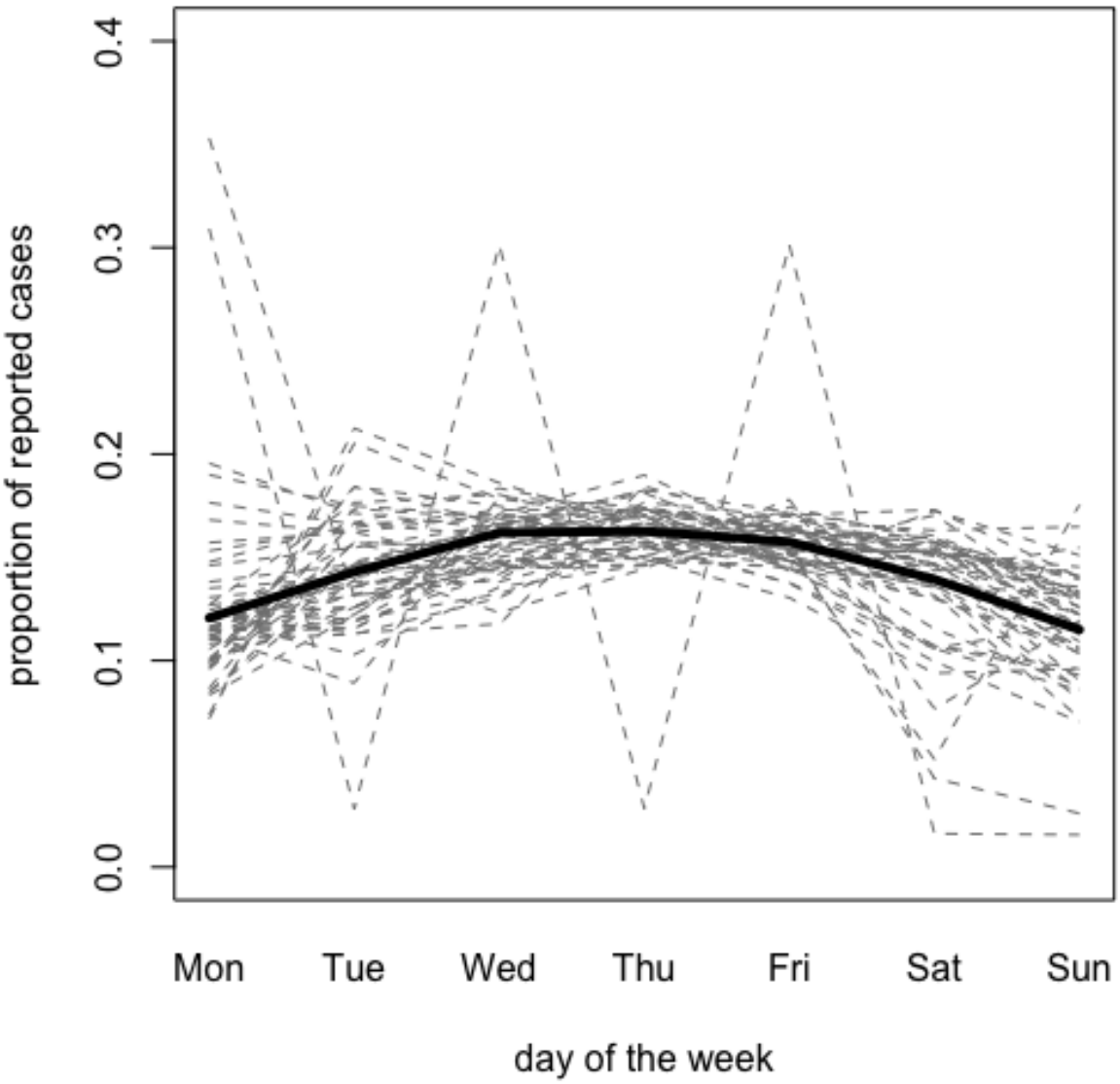

Figure 4.4: Proportion Of Case Counts On Each Day Of The Week And Mean For All 50 States

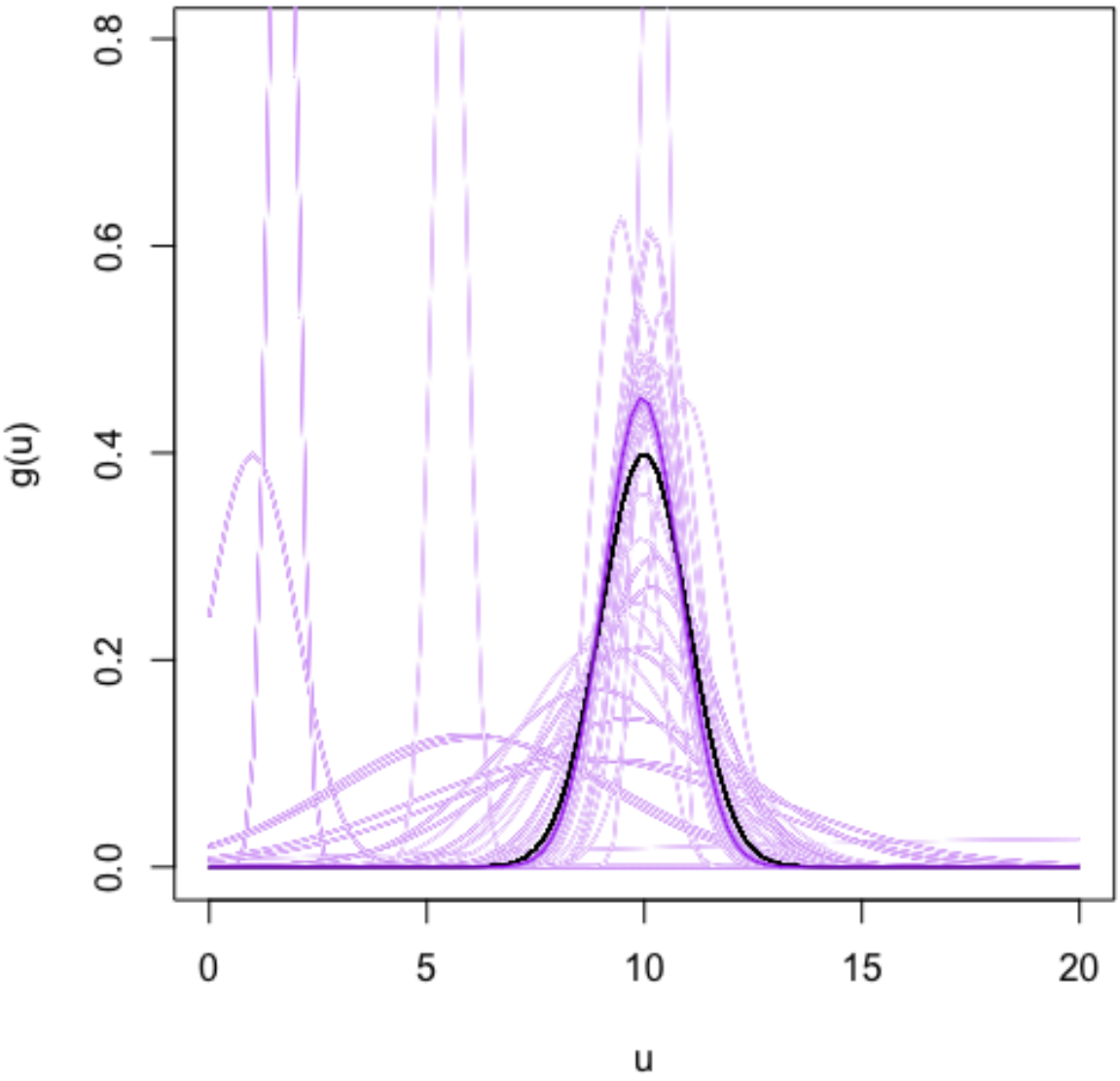

Figure 4.5: Estimated Normal Transmission Time Densities And Mean For All 50 States

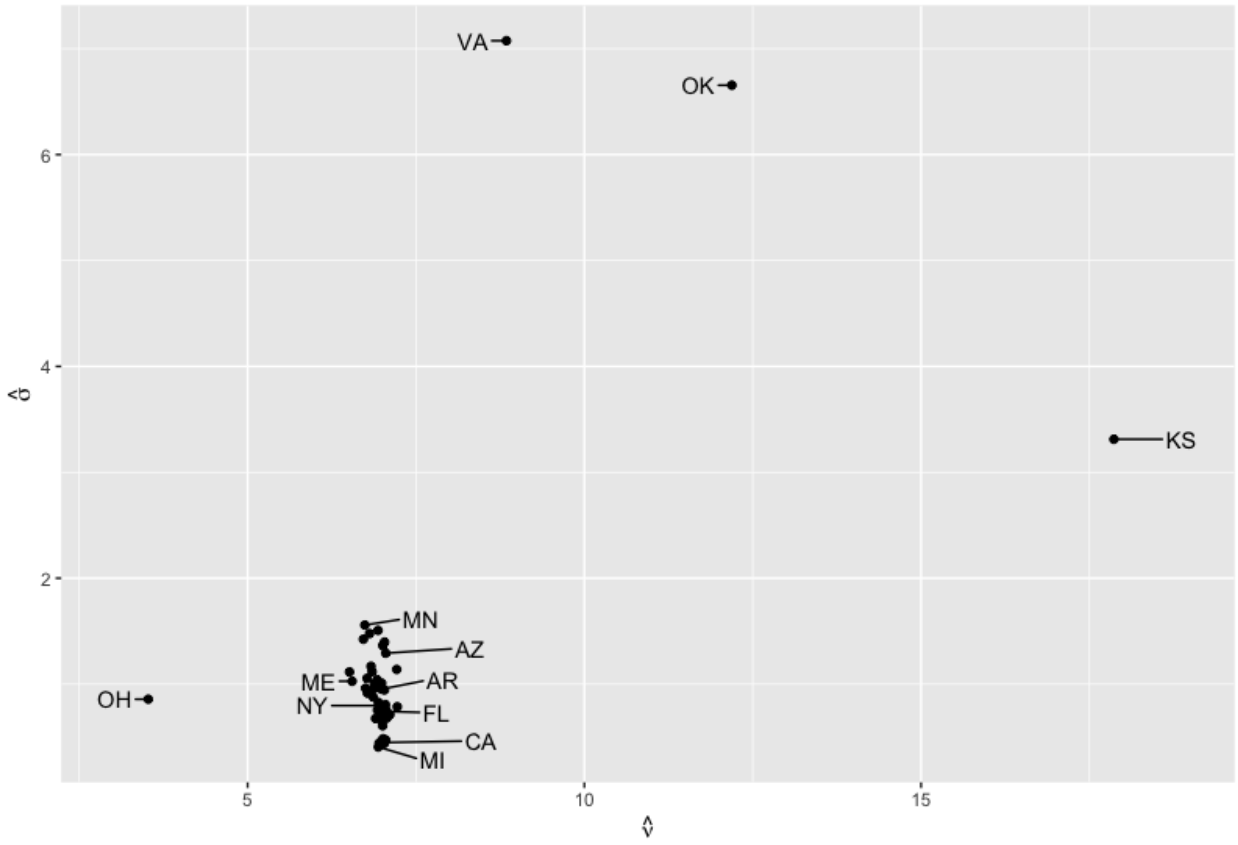

Figure 4.6: Estimates of the mean and standard deviation for estimated normal transmission time densities for fitted Hawkes models for all 50 states

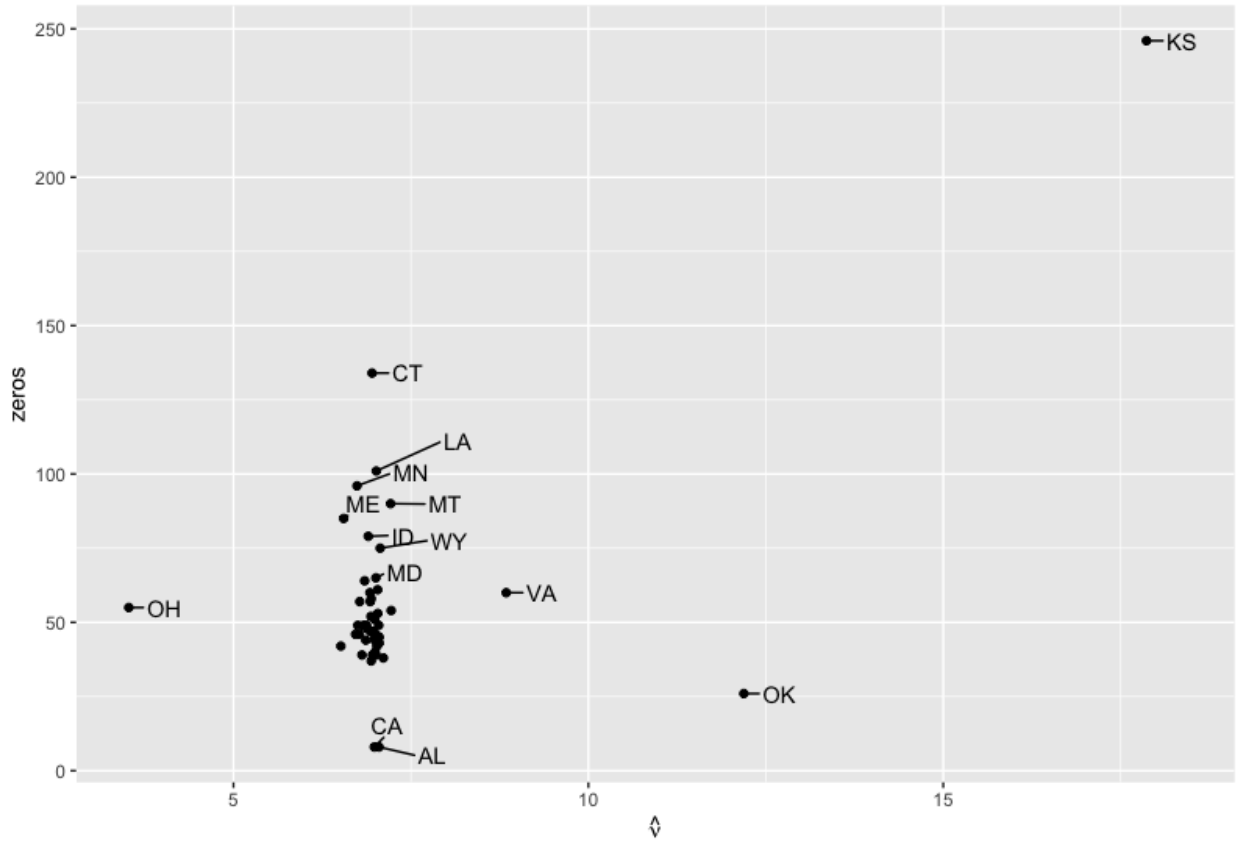

Figure 4.7: Estimates of the mean of the estimated normal transmission time density for fitted Hawkes models, for all 50 states, vs. the number of days in each state with zero confirmed SARS-CoV-2 cases in the dataset.

4.1 Discussion

The results suggest that the transmission time distribution for SARS-CoV-2 is centered at seven days, with a standard deviation of approximately one day. This is a somewhat narrower density compared to prior reports based on case studies.

The density in the estimate of the triggering function at 14 days is likely due to harmonic aliasing (Brillinger, 1981). However, the estimated density at one day is more difficult to explain. One possibility might be contagion due to physical contact, such as hand-to-hand or hand-surface-hand exposure. While most transmission is likely attributable to aerial spread via respiratory droplets, some cases might involve very short transmission times (Lotfi et al., 2020). Another explanation may be substantial autocorrelation between daily case counts, attributed to continuity in human behaviors, policies, and recording decisions. Additionally, some subjects might be infectious before their cases are reported, as recorded dates may differ from actual disease onset.

Data dumping and reporting trends might also explain some results. For instance, the higher estimated seven-day transmissions may result from weekly reporting cycles. However, as shown in Figure 4.2, transmission time estimates are not extremely sensitive to such errors, and most states' confirmed cases vary little by weekday. The results were remarkably consistent across states, with the seven-day density peaking.

It is important to note that many cases of SARS-CoV-2 are likely missing from state and CDC databases, particularly early in the pandemic when testing was scarce. While unreported asymptomatic cases were likely common, this underreporting seems minimally impactful on transmission time distribution estimates unless systematic trends exist. As discussed by Kresin, Schoenberg and Mohler (2021), the difficulty in estimating the percentage of asymptomatic cases introduces more error into SEIR models than Hawkes models, which bypasses this issue by focusing on recorded cases as an autoregressive process.

The assumption of a constant value of μ across states may be violated due to varying

immigration rates of the virus, especially as stay-at-home orders changed. However, such errors are unlikely to significantly affect transmission time estimates.

As an alternative to fitting a simple Hawkes model to each state, a multivariate Hawkes model could be used to account for interstate transmission. However, this approach may lead to issues of nonidentifiability and multicollinearity, resulting in high-variance estimates and poor forecasting performance (Yuan et al., 2021). Addressing these challenges is an important subject for future statistical work.

A crucial topic for future research is improving the estimation of uncertainties for Hawkes model parameters and forecasts. Current simulation-based methods tend to underestimate uncertainties, and exploring better ways to address this issue remains an important area for statistical work (Schoenberg, 2022).

Figure 4.8 considers a parametric form for the g function. Basically, on each day after 2020/01/23, we take the previous data and estimate μ, $K(t)$, c, p. Here, μ represents the baseline rate of new events, while K measures how strongly past events influence future ones. The parameter c denotes the average delay between a triggering event and subsequent cases, and p represents the variability of that delay. The distribution reveals interesting trends. At the very beginning, the range of different parameters is various covering 0 to 10. Gradually, the estimation of parameters converges to a specific range and becomes much more stable. μ concentrates within 0 to 4. $K(t)$ concentrates around 1. The mean of normal distribution concentrates within 8 to 10. The standard error concentrates around 0. With more data support, the estimation of parameters is more reliable and stable.

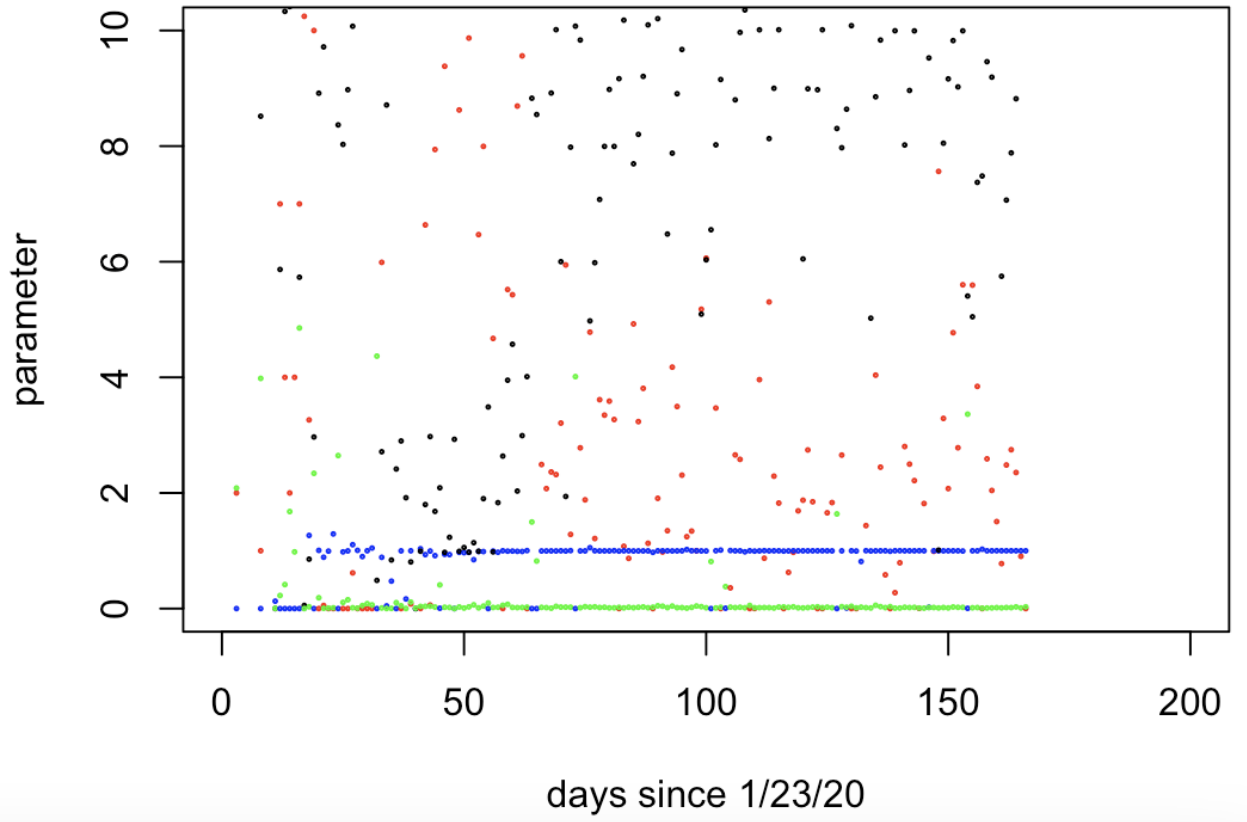

Figure 4.8: Model parameters estimation distribution over time. Each parameter is represented by a different color: red for μ, blue for K, black for c, and green for p.

CHAPTER 5

Conclusion

The work highlights the strength of Hawkes point process models in understanding the spread of SARS-CoV-2. By focusing on daily case counts, these models offer precise insights into transmission dynamics, outperforming traditional models like SEIR in terms of accuracy. The consistent finding of a seven-day transmission period suggests that the virus spreads faster and more predictably than earlier reports indicated, which has important implications for designing effective public health policies, such as quarantine durations.

The models also proved resilient in the face of real-world challenges, like missing data or irregular reporting, making them a reliable tool for pandemic analysis. However, the study recognizes that certain limitations remain. For instance, underreported cases, especially asymptomatic ones, could introduce some bias. Moreover, the assumption that the rate of external infections stays constant might not fully capture the complexities of how the virus enters different states.

Future research could build on this work by developing models that account for transmission between states or regions, as well as exploring better ways to handle uncertainties in the data. Despite these challenges, this study confirms that Hawkes models are a powerful and adaptable approach for analyzing infectious disease dynamics and can provide valuable support for managing public health responses.

CHAPTER 6

Supplementary Data and Reference

Frederic Schoenberg. (2024). *Estimating COVID-19 transmission time using Hawkes point processes.* ResearchGate.

Code for data analysis and simulations (DOI: 10.1214/23-AOAS1765SUPP; .zip).Zip file containing R code used for the data analysis, simulations, and construction.

REFERENCES

[1] World Health Organization. (2020). *Impact of COVID-19 on people's livelihoods, their health, and our food systems.*

[2] World Bank Blogs. (2020) *2020 Year in Review: The Impact of COVID-19 in 12 Charts.*

[3] World Health Organization (WHO). *Challenges to Healthcare Systems During COVID-19.*

[4] Science. (2020) *Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions.*

[5] American Medical Association. (2020). *Ensuring and sustaining a pandemic workforce.*

[6] Ali, S. T., Wang, L., Lau, E. H. Y., et al. (2020). Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science*, *369*(6507), 1106–1109.

[7] Anderson, R. M., & May, R. M. (1991). *Infectious diseases of humans: Dynamics and control*. Oxford University Press.

[8] Bacry, E., Mastromatteo, I., & Muzy, J. F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, *1*(1), 1550005.

[9] Bowsher, C. G. (2007). Modeling security market events in continuous time: Intensity-based, multivariate point process models. *Journal of Econometrics*, *141*(2), 876–912.

[10] Chernozhukov, V., Kasahara, H., & Schrimpf, P. (2021). Causal impact of masks, policies, behavior on early COVID-19 pandemic in the US. *Journal of Econometrics*, *220*(1), 23–62.

[11] Ferretti, L., Wymant, C., Kendall, M., et al. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, *368*(6491), eabb6936.

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

[13] Kim, J. Y., Choe, P. G., Oh, Y., et al. (2020). The impact of superspreading events on SARS-CoV-2 transmission. *Journal of Clinical Medicine*, *9*(9), 2986.

[14] Lemieux, J. E., Siddle, K. J., Shaw, B. M., et al. (2021). Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science*, *371*(6529), eabe3261.

[15] Mohler, G., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108.

[16] Nuzzo, J. B., Meyer, D., Snyder, M., & Cicero, A. (2020). What makes health systems resilient against infectious disease outbreaks and natural hazards? Results from a scoping review. *BMC Public Health*, *20*, 1310.

[17] Schoenberg, F. P. (2013). Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America*, *103*(1), 601–605.

[18] Zhuang, J., Ogata, Y., & Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research*, *109*(B5), B05301.

[19] CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC) (2021f). *COVID-19 Update for the United States, last accessed 9/14/21.*

[20] KRESIN, C., SCHOENBERG, F. and MOLER, G. (2021). *Comparison of Hawkes and SEIR models for the spread of Covid-19. Adv. Appl. Stat.*

[21] 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.*

[22] BAJEMA, K. L., WIEGAND, R. E., CUFFE, K., PATEL, S. V., IACHAN, R., LIM, T., LEE, A., MOYSE, D., HAVERS, F. et al. (2021). *Estimated SARS-CoV-2 seroprevalence in the US as of September 2020.*

[23] WERMER, E. and STEIN, J. (2020). Trump administration pushing to block new money for testing, trac-ing and CDC in upcoming coronavirus relief bill. *Washington Post*, 07/18/20.

[24] Rizodels, M. A., MISHRA, S., KONG, Q., CARMAN, M. and XIE, L. (2018). Linking epidemic models and Hawkes processes to model diffusions in finite population. In *Proceedings of the 2018 World Wide Web Conference* 419-428

[25] GUAN, W. J., Ni, Z. Y., HU, Y. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* 382 1708-1720.

[26] LAUER, S. A., GRANTZ, K. H., BI, Q., JONES, F. K., ZHENG, Q., MEREDITH, H. R., AZMAN, A. S., RE-ICH, N. G. and LESSLER, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* 172 577-582.

[27] LI, Q., GUAN, X., Wu, P., WANG, X., ZHOU, L., TONG, Y., REN, R., LEUNG, K. S. M., LAU, E. H. Y. et al.(2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* 382 1199-1207.

[28] HUANG, C., WANG, Y., LI, X., REN, L., ZHAO, J., Hu, Y., ZHANG, L., FAN, G., Xu, J. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395 497-506.

[29] WANG, D., Hu, B., Hu, C., ZHU, F., LIU, X., ZHANG, J., WANG, B., XIANG, H., CHENG, Z. et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China.*JAMA* 323 1061-1069.

[30] YANG, X., Yu, Y., XU, J., SHU, H., XIA, J., LIU, H., Wu, Y., ZHANG, L.., Yu, Z. et al. (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* 475-481.

[31] ZHOU, F., Yu, T., Du, R., FAN, G., LIU, Y., LIU, Z., XIANG, J., WANG, Y., SONG, B. et al. (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.

[32] CAUCHEMEZ, S., BOELLE, P. Y., DONNELLY, C. A., FERGUSON, N. M., THOMAS, G., LEUNG, G. M., HED- LEY, A. J., ANDERSON, R. M. and VALLERON, A. J. (2006). Real-time estimates in early detection of SARS. *Emerg. Infec. Dis.* 12 110.

[33] DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes: Elementary Theory and Methods. Vol. I, 2nd ed. Probability and Its Applications (New York).*Springer, New York. MR1950431

[34] REINHART, A. (2018). A review of self-exciting spatio-temporal point processes and their applications.*Statist. Sci.* 33 299-318. MR3843374

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

[36] OGATA, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes.*Ann. Inst. Statist. Math.* 30 243-261. MRO514494

[37] MARSAN, D. and LENGLINÉ, O. (2008). Extending earthquakes' reach through cascading. *Science* 319 1076-1079.

[38] MOHLER, G. (2013). Modeling and estimation of multi-source clustering in crime and security data.*Ann. Appl. Stat* 7 1525-1539. MR3127957

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

[40] KIRCHNER, M. (2016). Hawkes and INAR(oo) processes. *Stochastic Process. Appl.* 126 2494-2525. MR3505235

[41] KIRCHNER, M. (2017). An estimation procedure for the Hawkes process. *Quant. Finance* 17 571-593 MR3620953

[42] SCHOENBERG, F. (2023). Supplement to "Estimating Covid-19 transmission time using Hawkes point processes."

[43] BRILLINGER, D. R. (1981). *Time Series: Data Analysis and Theory, 2nd ed. Holden-Day Series in Time Series Analysis.* Holden-Day, Inc., Oakland, CA. MRO595684

[44] LOTEL, M., HAMBLIN, M. R. and REZAEL, M. (2020). COVID-19: Transmission, prevention, and potential therapeutic opportunities. *Clin. Chim. Acta* 508 254-266.

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

[46] SCHOENBERG, F. P. (2022). Nonparametric estimation of variable productivity Hawkes processes. Environ- metrics 33 Paper No. e2747, 13. MR4476429