Comparative Analysis of the SEIR and Point Process Models for Invasive Streptococcus Pneumoniae in Florida

Shu, Janella

UNIVERSITY OF CALIFORNIA

Los Angeles

Comparative Analysis of the SEIR and Point Process Models for Invasive *Streptococcus Pneumoniae* in Florida

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

by

Janella Shu

2021

ABSTRACT OF THE THESIS

Comparative Analysis of the SEIR and Point Process Models for Invasive *Streptococcus Pneumoniae* in Florida

by

Janella Shu
Master of Science in Applied Statistics
University of California, Los Angeles, 2021
Professor Frederic R. Paik Schoenberg, Chair

We investigated the extent to which a SEIR compartmental model, two Hawkes point process, each with a different trigger density function, and a recursive point process could characterize the transmission dynamics of invasive *Streptococcus pneumoniae.* All models were parameterized using surveillance data from Florida between 2010 to 2014. The maximum likelihood estimates of the parameters were calculated, and weekly counts were predicted using a thinning technique for the point processes and adaptive tau-leaping method for the SEIR model. Results suggest that the point processes performed better than the SEIR model. When comparing goodness of fit and prediction errors between the point processes, the recursive point process appeared to perform reasonably well on both. The recursive point process had an RMSE almost as small as the Hawkes with power-law decaying trigger density, which had the lowest RMSE, and the highest log-likelihood of all the models that were evaluated.

The thesis of Janella Shu is approved.

Yingnian Wu

Hongquan Xu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

In 1998, there were an estimated $62,840$ cases of invasive *Streptococcus pneumoniae* also known as invasive pneumococcal disease (IPD) in the United States (Bridy-Pappas, Margolis, Center, & Isaacman, 2005). *Streptococcus pneumoniae* is a leading cause of illness in young children, the elderly and persons with certain underlying medical conditions (Bridy-Pappas et al., 2005). In 2000, a 7-valent pneumococcal polysaccharide-protein conjugate vaccine (PCV7) was licensed for use among infants and young children (Advisory Committee on Immunization Practices, 2000). CDC's Advisory Committee on Immunization Practices (ACIP) recommends that the vaccine be used for all children aged 2-23 months and for children aged 24-59 months who are at increased risk for pneumococcal disease (Advisory Committee on Immunization Practices, 2000). CDC's Active Bacterial Core Surveillance (ABC) data indicated that in 2008 before a 13-valent pneumococcal conjugate vaccine (PCV13) replaced PCV7 for routine use among children, approximately 61% of IPD among children, younger than 5 years were attributable to the serotypes included in PCV13, and PCV7 serotypes caused less than 2% of cases (Hamborsky & Kroger, 2015). Indirect effects from PCV13 use among children might further reduce the remaining burden of adult pneumococcal disease caused by PCV13-types (Hamborsky & Kroger, 2015). In 2010, PCV13 was licensed and the Food and Drug Administration approved PCV13 as a single dose for people 50 years of age and older in December 2011 (Hamborsky & Kroger, 2015).

Since the introduction of PCV7 and PCV13, IPD has decreased from 100 cases per 100,000 people in 1998 to 9 cases per 100,000 in 2015 (*Pneumococcal Disease*, 2017). There is evidence to support that serotype replacement, an increase in the incidence of IPD caused by nonvacine types (NVTs) after vaccine introduction, had occurred in IPD after the introduction of PCV7 in most populations and was caused by the vaccine (Weinbergera, Malley, & Lipsitch, 2012). There are at least 93 known serotypes of pneumococci. Therefore, it is not unreasonable to assume that serotype replacement will eventually occur after the introduction of PCV13 as well. Analyzing the infectious disease dynamics can help inform decisions on prevention and control of IPD.

Although the number of cases in the United States are very low, countries in sub-Saharan Africa continue to have large outbreaks including in December 2015 and April 2016 (Organization, 2016). In Malawi, a landlocked country in southern Africa that lies to the south of the classical meningitis belt, *Streptococcus pneumoniae* is a leading cause of pneumonia, sepsis, bacterial meningitis and death in both children and adults (Everett & et al. ., 2011). The high antimicrobial resistance of *Streptococcus pneumoniae* in Asia contributes to both the treatment and economic burden caused by IPD (Bravo, 2009). The Asian Strategic Alliance for Pneumococcal Disease Prevention (ASAP) is a work group composed of healthcare professionals from 12 Asian countries and territories; namely, Hong Kong, India, Indonesia, Korea, Macau, Malaysia, Pakistan, the Philippines, Singapore, Sri Lanka, Taiwan and Thailand (Bravo, 2009). Pneumococcal vaccination is currently not part of the national immunization programs of any of the ASAP member countries and territories, despite it being projected to prevent around $260,000$ deaths annually as well as having the potential to mitigate widespread antibiotic resistance (Bravo, 2009).

The objective of a mathematical model of an infectious disease is to describe the transmission process of the disease (Li, 2018). There are three general approaches to mathematical modeling of infectious diseases: statistical models, deterministic models, and stochastic models (Li, 2018). In this paper, we will be comparing a deterministic SEIR compartmental model, and three stochastic models, specifically Hawkes process with exponentially decaying trigger density, Hawkes process with power-law decaying trigger density and a point process proposed by Schoenberg, Hoffmann and Harrigan (2019) called the recursive point process. A Hawkes process is self-exciting but also has the property of static productivity. However, this assumption of static productivity seems questionable (Schoenberg, Hoffmann, & Harrigan, 2019). Early in the onset of an epidemic, when prevalence of the disease is still low, one would expect the rate of transmission to be much higher than when the prevalence of the disease is higher, because of human efforts at containment and intervention of the disease, and because some potential hosts of the disease may have already been exposed (Schoenberg et al., 2019).

The organization of the paper is as follows: Chapter 2 defines a Hawkes and recursive point process model. Chapter 3 defines a SEIR model. In Chapter 4, we describe the data, the methods used to estimate model parameters as well as goodness of fit measures. In Chapter 5, we provide the results from our analysis including the estimated parameters, model performance and weekly predictions. Finally in Chapter 6, we discuss possible improvements and alternative models.

CHAPTER 2

Point Process Model

2.1 Point Process

A (simple) point process is a sequence of random variables $\boldsymbol{T} = t_1, t_2, ...$ taking values in $[0, \infty)$, has $\mathbb{P}(0 \geq t_1 \geq t_2 \geq ...) = 1$ and the number of points in the bounded region is almost surely (a.s.) finite (Laub, Taimre, & Pollett, 2015). For a temporal point process, t_i would represent the times of events occurring between time 0 and t (Schoenberg, 2010). A temporal point process, N, can be alternatively described as a count process $N(t)$. A count process is a stochastic process $(N(t) : t \geq 0)$ taking values $\mathbb{N}_0$ that satisfies $N(0) = 0$, is a.s. finite, and is a right-continuous step function with increments of size +1 (Laub et al., 2015; Daley & Vere-Jones, 2003).

The behavior of a simple temporal point process N is typically modeled by specifying its conditional intensity $\lambda(t)$ (Schoenberg, 2010).

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

where $\mathcal{H}(t)$ is the history of arrivals up to time t.

2.2 Hawkes

Alan G. Hawkes described a class of point processes that were self-exciting, where the current intensity of events is determined by events in the past (Hawkes, 1971). He suggested that the self-existing process is a possible epidemic model in large populations in so far as the occurrence of a number of cases increases the probability of further cases (Hawkes, 1971). The Hawkes process is a mathematical model for these self-exciting processes. For a simple

Hawkes process, the conditional rate of events at time t can be written as

$$\lambda(t) = \mu + K \int_0^t g(t-t')dN(t') = \mu + K \sum_{i:t_i<t} g(t-t_i)$$

where $\mu > 0$, is the background rate, $g(v) \geq 0$ is the trigger density satisfying $\int_0^\infty g(u)du = 1$ which describes the conductivity of events, and the constant K is the productivity, defined as the expected number of first generation offspring of the point t_i, which is required to satisfy $0 \leq K < 1$ in order to ensure stationary and subcriticality (Schoenberg et al., 2019). A frequently used trigger function is the exponential decay function, $\alpha e^{\beta(t)}$ where α, $\beta > 0$, which can be interpreted as each arrival in the system instantaneously increases the arrival intensity by α and over time this arrival's influence decays at rate β (Laub et al., 2015). Another commonly used function is the power-law function, $\frac{k}{(c+(t-t'))^p}$ where c and p are positive scalars, which was popularized by the geological model called Omori's Law that was used to predict the rate of aftershocks caused by an earthquake (Laub et al., 2015).

Some applications of the Hawkes point process include using high frequency financial data to model the so-called volatility phenomenon at the transactional level, estimating residential burglary and gang violence to create crime hotspot maps, and characterizing the temporal pattern of seismicity (Bacry, Mastromatteo, & Muzy, 2015; Mohler, Short, Brantingham, Schoenberg, & Tita, 2011; Ogata, 1998).

2.3 Recursive

Schoenberg et al. (2019) proposes an extension of the Hawkes point process that is recursive. In particular, the productivity of this model at time t is a function of the conditional intensity at t and the conditional intensity in turn depends critically on this productivity (Schoenberg et al., 2019) . The model can be written as

$$\lambda(t) = \mu + H(\lambda_{t'})g(t-t')dN(t')$$

where $\mu > 0$, $g > 0$ is the density function, $\lambda_{t'}$ means $\lambda(t')$ and H is the productivity function which would typically be a decreasing function (Schoenberg et al., 2019). In our analysis, we use $g(u) = \beta e^{-\beta u}$. Suppose $H(x) = \kappa x^{-\alpha}$ where $\kappa > 0$ then,

$$\lambda(t) = \mu + \kappa \int_0^t \lambda_{t'}^{-\alpha} g(t - t') dN(t').$$

When $\alpha = 0$, $\lambda(t)$ is a Hawkes point process. Schoenberg (2019) refers to $\lambda(t)$ where $\alpha = 1$ as the standard recursive model. For the Hawkes model, the productivity at each point t_i is K, but for a standard recursive model, the average productivity is $\frac{\kappa T}{N(S)} \to \frac{\kappa}{\mu+\kappa}$ a.s. since $\frac{N(S)}{T} \to \mu + \kappa$ a.s., where $S = [0, T]$ (Schoenberg et al., 2019).

CHAPTER 3

Compartmental Epidemic Model: SEIR

Deterministic or compartmental models are one approach to describing the transmission process of a disease. These are typically models using differential and difference equations of various forms. The host population is partitioned into mutually exclusive groups (i.e. compartments) according to the natural history of the disease. For a simple infectious disease, possible compartments include susceptible (S), infected (I) and recovered (R). These models describe the dynamic interrelations among the rates of change and population sizes (Li, 2018).

In 1906, William Hamer suggested that the course of an epidemic depends on the rate of contact between susceptible and infectious individuals, and in 1908 Ronald Ross translated Hamer's discrete-time model into a continuous time framework (Anderson, 1991). The ideas of Hamer and Ross were explored in more detail by William Kermack and Anderson McKendrick who proposed a model that was compartmental and deterministic in structure and where the total population was assumed to be constant and divided into three compartments: susceptible (S), infected (I) and removed/recovered (R).

The Kermack–McKendrick model is a particular instance of the SIR model. Suppose an infectious disease has a latent period, one way to incorporate the disease latency in a mathematical model is to split the infected compartment into an exposed/latent compartment (E) and an infected compartment (I). This is called a SEIR model and is an extension of the SIR model. Since IPD persists for a short period of time, we assume a closed population with no births or deaths. The SEIR model can be described by the following equations:

$$\frac{dS}{dt} = -\beta(t)\frac{SI}{N}$$
$$\frac{dE}{dt} = \beta(t)\frac{SI}{N} - \sigma E$$
$$\frac{dI}{dt} = \sigma E - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

where the total population, N, is the sum of all individuals in S,E, I and R, and $\beta(t)$ is the time dependent transmission rate. In our analysis, we assume that the transmission rate decays exponentially over time, $\beta(t) = \beta_0 e^{-kt}$ and that σ, the rate of death/recovery, and γ, the rate of latent individuals becoming infectious, are constant. Therefore, $1/\sigma$ and $1/\gamma$ are the mean latent and infectious period, respectively. Figure 3.1 shows the transfer diagram of the SEIR model.

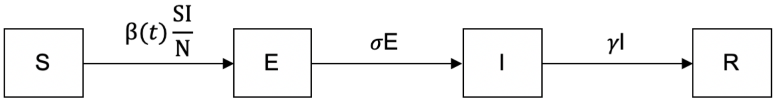


Figure 3.1: Transfer diagram of SEIR model with constant population

The basic reproductive number, R_0 is the single most important parameter in epidemic modeling (Li, 2018). It measures the average number of secondary infections caused by a single infectious individual in an entirely susceptible population during the mean infectious period (Li, 2018). For the SEIR model in Figure 3.1, $R_0(t) = \beta/\gamma$. If $R_0 < 1$ then an epidemic will not occur, and if $R_0 > 1$ then an epidemic will occur (Li, 2018).

Deterministic models have been used to describe a variety of diseases. For example, Roy Anderson (1988) developed simple deterministic models in order to address two problems in the study of the transmission dynamics of HIV-1: the variability in incubation and infectious periods, and heterogeneity in sexual activity within homosexual and heterosexual communities in the United Kingdom. The basic and effective reproduction numbers of the

2014 Ebola virus (EBOV) outbreak in West Africa have been estimated using a SEIR model to better understand the spread of infection in the affected countries (Althaus, 2014). That study also provided real-time estimates of EBOV transmission parameters during the ongoing outbreak. Additionally, another study examined mathematically the impact of isolation and quarantine on the control of SARS during the outbreaks in Toronto, Hong Kong, Singapore and Beijing using a deterministic model containing six compartments: susceptible, asymptomatic, quarantined, symptomatic, isolated and recovered (Gumel et al., 2004). The model reasonably mimics the outbreaks in geographically distinct regions, supporting the notion that simple models can be used to provide insights into the dynamics and control of an epidemic in progress (Gumel et al., 2004). The study showed that the size and duration of an outbreak can be greatly influenced by the timely implementation of the isolation program (Gumel et al., 2004).

CHAPTER 4

Model Development

4.1 Description of Data

Project Tycho was created in 2013 to improve access to standardized data in global health (van Panhuis, Cross, & Burke, 2018). The first version of Project Tycho (v1) was comprised of over a century of infectious disease surveillance data from 1888 to 2014 for the United States. Version 2 (v2) has been updated with new weekly US surveillance data in addition to surveillance data for dengue-related conditions from 98 additional countries (van Panhuis et al., 2018). All Project Tycho data are represented as counts of cases or deaths to disease conditions reported by public health surveillance, and SNOMED-CT terminology is used to represent the reporting of cases or deaths due to diseases, and diagnostic certainty (van Panhuis et al., 2018).

We obtained the pre-compile IPD dataset from the Project Tycho website (`http://www.tycho.pitt.edu`) which contained 42,208 records of weekly case count from 12/29/2002 to 12/30/2017. There are 20 variables in the dataset, including the start and end date of a period, the number of cases in that period, whether the count is part of a cumulative count series or not, and the age range and location (by state) for those cases. Since there is missing data from 01/01/2003 to 01/01/2004, 11/04/2007 to 12/31/2009, and 08/03/2014 to 12/31/2014, we only included cases that occurred between 01/01/2010 to 08/02/2014 in our analysis. We chose to focus on IPD cases in Florida because it had more consistent weekly case counts in our chosen date range. As in Park et al. (2018), estimated occurrence times were distributed uniformly within report date ranges. The training and testing data sets were created by a 90:10 split. Figure 4.1 is a plot of the weekly counts for the training (blue dots) and test (red dots) set.

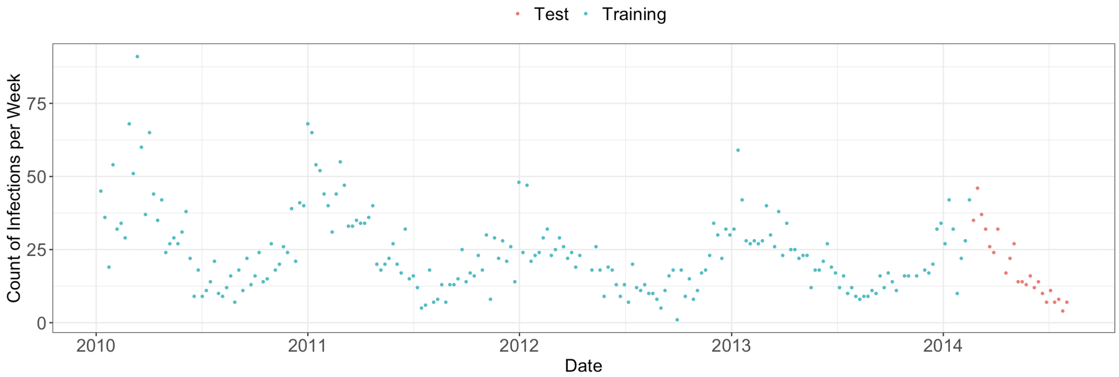


Figure 4.1: Weekly counts of infection in Florida from 01/01/2010 to 08/02/2014

4.2 Parameter Estimation

Given occurrence observation $t_1, t_2, ..., t_n$ for an interval $[0, T]$, the log-likelihood of a point process with an intensity function

$$\lambda(t|\theta) = \mu + \int_{-\infty}^{t} g(t-u|\theta)dN(u)$$

is given by

$$log\,\mathcal{L}(t_1, ..., t_n|\theta) = -\int_0^T \lambda(t|\theta)dt + \int_0^T log\lambda(t|\theta)dN(t)$$

where $\theta = (\theta_1, \theta_2, ..., \theta_r)$ (Ozaki, 1979). As the log-likelihood is non-linear with respects to the parameters, the maximum of the log-likelihood is performed by using non-linear optimization techniques. One of the optimization techniques that Ozaki (1979) mentions is a direct method. The direct method that we used to estimate the parameters for the Hawkes model and recursive model is the Nelder-Mead simplex method (Nelder & Mead, 1965).

For the SEIR model, maximum likelihood estimates (MLE) of the parameters were obtained by fitting the model to the data, assuming that cumulative numbers of cases are Poisson distributed (Althaus, 2014). The Nelder-Mead simplex method was also used to estimate the parameters for the SEIR model.

4.3 Model Assessment

The goodness of fit of all models was assessed using the method of super-thinning proposed by Clements, Schoenberg and Veen (2012) to perform residual analysis. For the SEIR model, we used $\beta(t)$ multiplied by the infectious population at time t as an estimate of the conditional intensity function $\hat{\lambda}(t)$ (Park, Chaffee, Harrigan, & Schoenberg, 2020). Suppose N is a temporal point process with conditional intensity $\lambda(t)$ then to transform N into a residual Poisson process with rate k where $inf\{\hat{\lambda}(t)\} \leq k \leq sup\{\hat{\lambda}(t)\}$, first thin N keeping each point t_i independently with probability $min\{\frac{k}{\hat{\lambda}(t_i)}, 1\}$ to obtain a thinned residual process Z_1 (Clements, Schoenberg, & Veen, 2012). Next we simulated a homogeneous Poisson process with rate k and independently keeping each simulated point t_j with probability $max\{\frac{k-\hat{\lambda}(t)}{k}, 0\}$ to obtain a Cox process Z_2 (Clements et al., 2012). The points of the residual point process $Z = Z_1 + Z_2$ are called super-thinned residual points because Z is homogeneous Poisson with rate k if and only if $\hat{\lambda} = \lambda$ almost everywhere (Clements et al., 2012). Since the tuning parameter k controls the rate of thinning and superposition, we chose k to be the mean of estimated lambdas, $\hat{\lambda}(t_i)$ where $t_i \in T$, since it minimizes the sum of squared deviations of the estimated conditional intensity from k (Clements et al., 2012).

For the point processes, we calculated the Stoyan-Grabarnik diagnostic. Baddeley (2005) noted that for a spatial point process $\mathbf{x} = \{x_1, ..., x_n\}$,

$$\mathbb{E}\left[\sum_{x_i \in \mathbf{X} \cap B} \frac{1}{\lambda(x_i, \mathbf{X})}\right] = \mathbb{E}\left[\int_B 1du\right] = |B|$$

where $|B|$ denotes the area of B (Baddeley, Turner, Møller, & Hazelton, 2005). If each point x_i of $\mathbf{X}$ is weighted by the reciprocal of its Papangelou conditional intensity $m_i = \frac{1}{\lambda(x_i, \mathbf{X})}$, called the "exponential energy mark" by Stoyan and Grabarnik, then the total weight of all points x_i in $\mathbf{X}$ fall in a nominated region B,

$$M(B) = \sum_{x_i \in \mathbf{X} \cap B} m_i$$

has expectation $\mathbb{E}[M(B)] = |B|$ (Baddeley et al., 2005). Similarly, for a temporal point process, instead of the Papangelou conditional intensity, we used the conditional intensity and since points of a Poisson process are independent, $m_i = \frac{1}{\lambda(x_i)}$ and for a temporal Poisson process, B is a temporal window . Therefore, if $\hat{\lambda}$ is a good estimate, then

$$\frac{\sum\limits_{t_i \in T} \frac{1}{\hat{\lambda}(t_i)}}{T}$$

should be approximately 1.

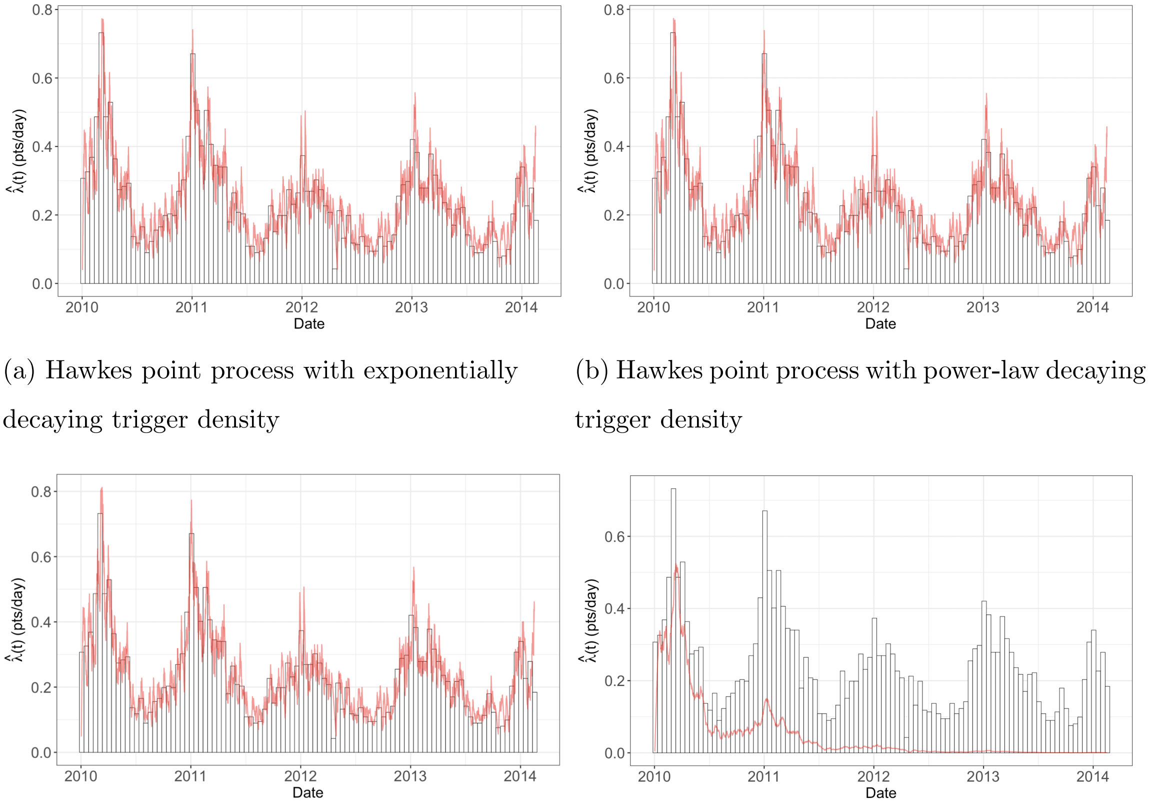
Additionally, for all the models, we determined the Akaike Information Criterion (AIC) which is defined as $2k - 2ln(L)$, where k is the number of estimated parameters and L is the maximum value of the likelihood function of the model. AIC was used to compare the models' goodness of fit relative to each other.

CHAPTER 5

Results

The estimated parameters for the Hawkes process with exponentially decaying trigger density, $g(u) = \beta e^{-\beta u}$, are $(\hat{\mu},\ \hat{K},\ \hat{\beta}) = (0.5490482,\ 0.8373538,\ 0.3152318)$ with corresponding standard error estimates (0.08155830, 0.02676473, 0.03019345). Since $g(u)$ is the exponential density function, $\frac{1}{\hat{\beta}} = 3.17$ days can be interpreted as the estimated mean delay time between events that excites each other. The estimated parameters for the Hawkes process with power-law decaying trigger density, $g(u) = \frac{(p-1)c^{(p-1)}}{(u+c)^p}$, are $(\hat{\mu},\ \hat{K},\ \hat{c},\ \hat{p}) =$ (0.5242630, 0.8425341, 41.3548356, 14.1236510) with corresponding standard error estimates (0.08241387, 0.02705487, 55.38658258, 17.05229724). From the standard error estimates, we can see that $\hat{c}$ and $\hat{p}$ have very large variance compared to the other estimated parameters. In both Hawkes point processes, on average, 0.84 cases are triggered by any given infected individual and both models have an estimated background rate of infection of approximately 1 case every 2 days. The estimated parameters for the recursive point process with exponentially decaying trigger density, $g(u) = \beta e^{-\beta u}$, are $(\hat{\mu},\ \hat{k},\ \hat{\beta},\ \hat{\alpha}) =$ $(0.67385384,\ 0.70639551,\ 0.32003470,\ -0.09174118)$ with corresponding standard error estimates (0.13298941, 0.09696835, 0.03206618, 0.06916675). The small value of $\hat{\alpha}$ seems to suggest that IPD may have a constant productivity and therefore the recursive model could be reduced to a Hawkes point process.The estimated parameters of the SEIR model are $(\hat{\beta}_0,\ \hat{k},\ \hat{\sigma},\ \hat{\gamma}) = (0.054768393,\ 0.003812852,\ 0.333333333,\ 0.058823529)$. We fixed $\hat{\sigma}$ and $\hat{\gamma}$ to be $\frac{1}{3}$ and $\frac{1}{17}$ since we assumed that the latent and infectious period is 3 days and 17 days, respectively (Melegaro et al., 2010; Hamborsky & Kroger, 2015). Additionally, we set N equal to $19,150,000$ (Bureau of Economic and Business Research & University of Florida, 2014). The assumed latent period of the SEIR model, 3 days, has a similar interpretation as $\frac{1}{\hat{\beta}} = 3.17$

days from the Hawkes model with exponentially decaying trigger density and has a similar estimated value as well. Since we assumed the transmission rate to decay exponentially, it is a monotonically decreasing function and therefore, the maximum of $\beta(t)$ is β_0. The basic reproductive number $R_0(t) = \frac{\beta(t)}{\gamma} \leq \frac{\beta_0}{\gamma} = 0.931$ suggests that the initial case will not lead to an outbreak. R_0 can be interpreted as the expected number of secondary cases in a completely susceptible population (Delamater, Street, Leslie, Yang, & Jacobsen, 2019). The population of Florida is not a complete susceptible population, but as previously mentioned, PCV7 serotype caused less than 2% of cases in children younger than 5 years old in 2008. So, one might assume that Florida was susceptible to the serotypes not covered in PCV7 but was eventually covered by PCV13.



(a) Hawkes point process with exponentially decaying trigger density

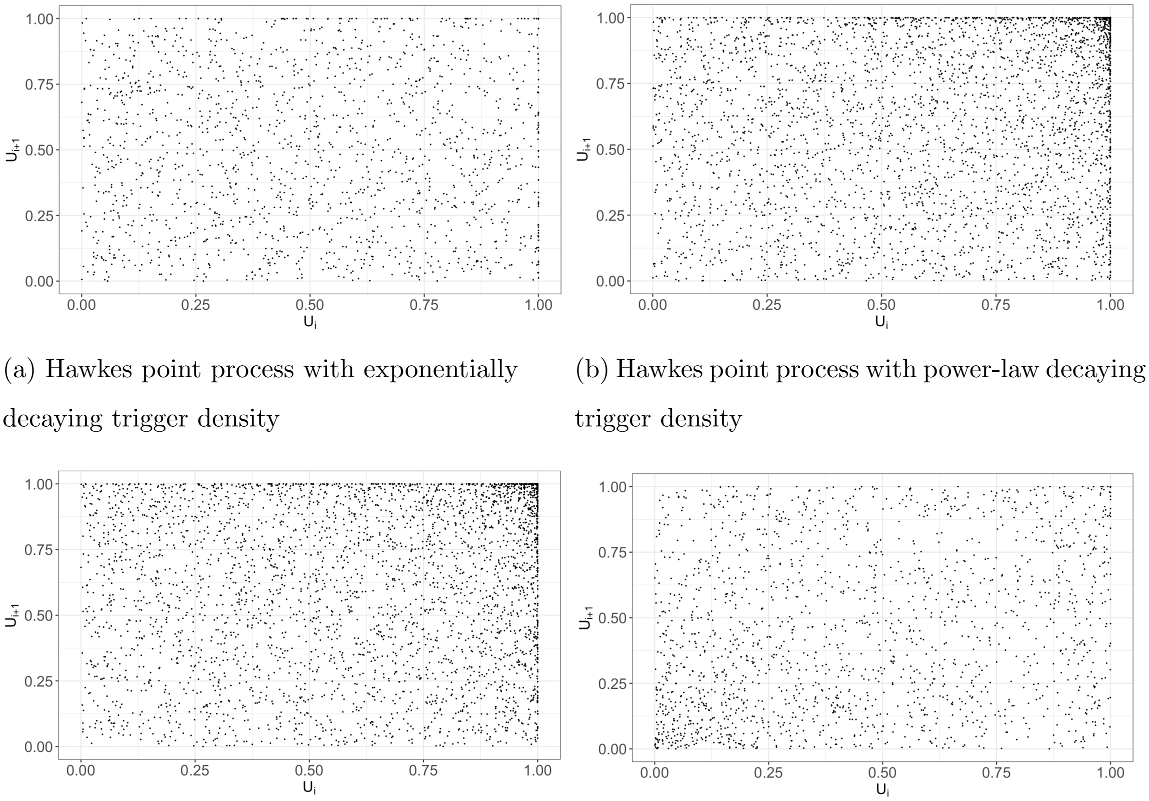
(b) Hawkes point process with power-law decaying trigger density

(c) Recursive point process

(d) SEIR

Figure 5.1: Histogram of IPD in Florida from 01/01/2010 to 02/15/2014 with fitted intensity (red)

In Figure 5.1, we can see that the estimated conditional intensity of the SEIR model does not fit the data as well as the estimated conditional intensity of the point process models. To calculate the estimated conditional intensity, $\hat{\lambda}(t)$, for the SEIR model, we multiplied the infectious population at time t_i by $\hat{\beta}(t_i) = \hat{\beta}_0 e^{-\hat{k}t_i}$. The infectious population at time t_i, was assumed to be the total number of cases observed during the 17 days, the assumed infectious period, prior to time t_i. It is also evident that the recursive model has a higher $\hat{\lambda}$ at the peak near 2011.



(a) Hawkes point process with exponentially decaying trigger density

(b) Hawkes point process with power-law decaying trigger density

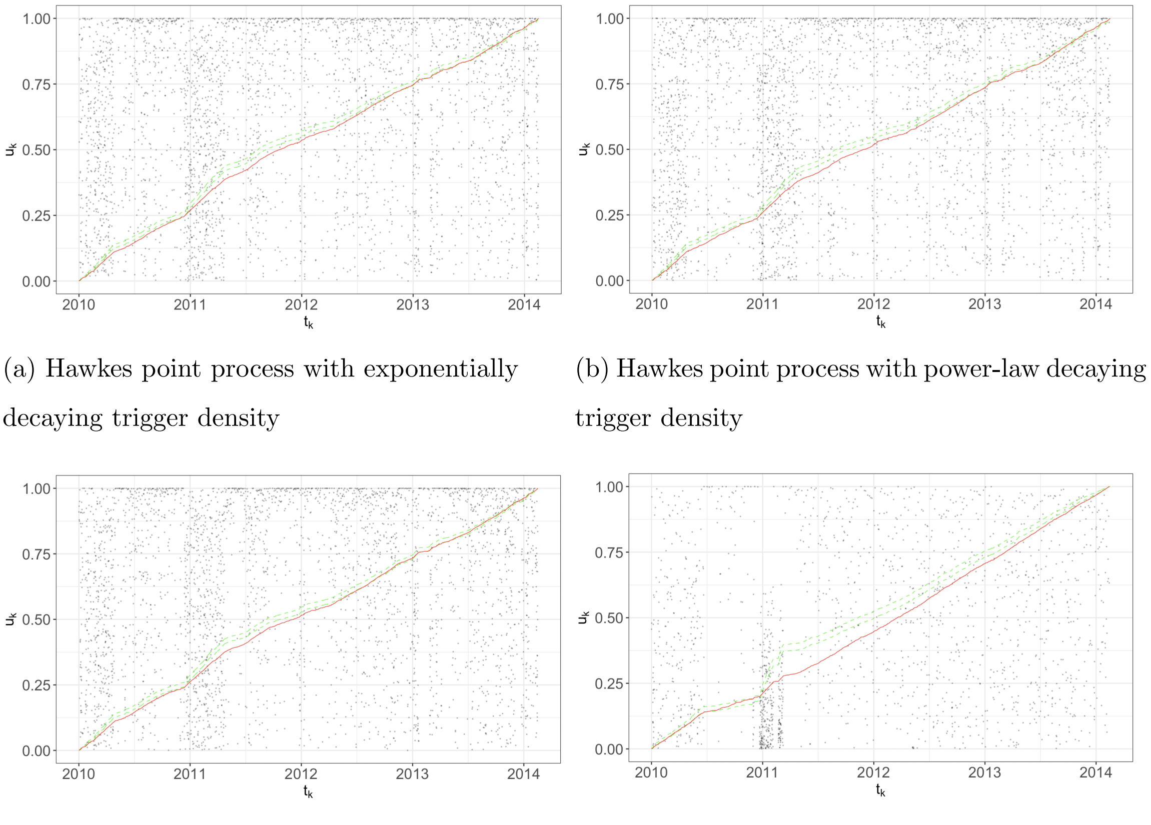
(c) Recursive point process

(d) SEIR

Figure 5.2: Lag plot of the standardized interevent times u_i of the super-thinned residuals

The lag plots of the standardized interevent times of the super-thinned residuals u_i where b is the mean of $\hat{\lambda}(t_i)$ are shown in Figure 5.2. Overall, the points on all the lag plots look relatively well scattered. For the Hawkes with power-law decaying trigger density, and recursive model there seems to be clustering in the upper right corner and for the SEIR model, there the points seem to be more concentrated at the lower left corner.

Figure 5.3 shows the super-thinned residuals t_i along with their corresponding standardized interevent times u_i, the cumulative sum of the standardized interevent times (solid red line), and the individual 95% confidence bounds based on 1,000 simulations of an equivalent number of uniform random variables (dashed green line). Again, we chose b to be the mean of $\hat{\lambda}(t_i)$. In all four plots, there seems to be clustering of points around the first third of 2010 and 2011 and more sparsity during the remaining part of those 2 years. This indicates that the models underestimate the first 4 months of 2010 and 2011 and overestimates the last 8 months of 2010 and 2011. For 2013 and 2014, there seems to be less of a discernible pattern.



(a) Hawkes point process with exponentially decaying trigger density

(b) Hawkes point process with power-law decaying trigger density

(c) Recursive point process

(d) SEIR

Figure 5.3: Super-thinned residuals t_k and their corresponding standardized interevent times u_k

The Stoyan-Grabarnik diagnostics, log-likelihood and AIC are shown in Table 5.1. For all point processes, the Stoyan-Grabarnik diagnostics is very close to 1, which indicates that $\hat{\lambda}$ is a good estimate. Since the recursive and Hawkes with exponentially decaying trigger density models are nested, we can calculate a likelihood ratio test where $LRT = -2(ln(\mathcal{L}_s(\hat{\theta})) - ln(\mathcal{L}_g(\hat{\theta})))$ and s is the simpler model and g is the general model. The difference in log-likelihood is approximately chi-squared with degrees of freedom equal to the difference in the number of parameters between the two models. Therefore, with a chi-square score of 3.888 and $df = 4 - 3 = 1$, the improvement in fit by the recursive mode is statistically significant with significance level of 0.05. The recursive model has the lowest AIC which would suggest that it is the better model, but differences in AIC between the recursive and the other two Hawkes models are marginal when compared to the difference between the recursive and the SEIR model. The recursive model also has the largest log-likelihood.

	Hawkes, exponential	**Hawkes, power-law**	**Recursive**	**SEIR**
Goodness of fit				
Stoyan-Grabarnik	0.9999948	1.002954	1.000396	n/a
Log-likelihood	1660.985	1660.928	1662.929	-4021.85
AIC	-3315.97	-3313.856	-3317.892	8047.7
Weekly Prediction Results				
RMSE from prediction	7.389659	7.259191	7.280142	10.8545
RMSE from overpredicting	7.57113	6.971462	7.173118	8.41517
RMSE from underpredicting	6.929185	8.06099	7.533728	11.15954
SSE	1310.569	1264.7	1272.011	2827.685
SSE% from overpredicting	74.36%	69.17%	68.77%	7.51%
SSE% from underpredicting	25.64%	30.83%	31.23%	92.49%

Table 5.1: Stoyan-Grabarnik diagnostic, log-likelihood, AIC and squared error from model fitting

To create predictions for the point processes, we simulated a one-dimensional nonhomogeneous Poisson process using an algorithm described by Lewis and Shedler (1979), which simulated a nonhomogeneous Poisson process with rate $\lambda(x)$ by generating points from nonhomogeneous Poisson process, $\{N^*(x) \geq 0\}$, with rate $\lambda^*(x)$ on a fixed interval $(0, x_0]$ where $\lambda^*(x) \geq \lambda(x)$ and then keeping the simulated points with probability $\frac{\lambda(t_i)}{\lambda^*(t_i)}$. For our simulation, we chose λ^* to be 14 which is greater than the maximum of $\hat{\lambda}(t_i)$. To simulate weekly counts for the SEIR model, we used the Gillespie's algorithm with a tau-leaping approximation which was introduced by Cao (2007) and has been implemented for a compartmental epidemic model (Rivers, Lofgren, Marathe, Eubank, & Lewis, 2014). For the Gillespie algorithm, first, draw a randomly exponentially-distributed waiting time until the next transition and then randomly select the identity of the transition, weighted by the relative transition rates (Johnson, 2014). When models have larger transition rates an approximation approach is necessary to increase simulation speed (Johnson, 2014). One way to perform this approximation is the adaptive tau-leaping algorithm to reduce the number of iterations by treating transition rates as constant over time for which this approximation leads to little error (Johnson, 2014). To implement the tau-leaping algorithm to predict weekly counts, we set the initial infectious population as the number of cases observed 17 days (duration of infectious period) prior to the predicted week, the initial susceptible population is the total population of Florida minus the initial infectious population, and the initial exposed and recovered population is zero.

We re-estimated the model's parameter using all the data prior to the week that we were predicting. Additionally we did 1,000 simulations for each week and took the average to be the predicted weekly count. Figure 5.4 shows the predicted weekly counts for each model and the actual count (black line). SEIR underpredicted cases for all weeks with the exception of the 8th week, where there were zero cases, and the 19th week. The prediction for the SEIR model seem to improve in the 12th week and in subsequent weeks. All the point process models have similar weekly predictions. In the last 12 weeks it looks like the Hawkes with exponentially decaying trigger density had slightly higher predictions than the Hawkes with power-law decaying trigger density and recursive model. The predictions of the Hawkes with power-law decaying trigger density (blue line) and recursive models (yellow-green line) are very similar for a majority of the forecasted weeks. No model predicted the zero cases in the 8th week or the peaks at 2nd, 7th and 11th week. RMSE is reduced by approximately 32.93% when we compare the RMSE of the SEIR model to the recursive model in Table 5.1. The Hawkes models have a similar level of RMSE reduction. The difference between the RMSE and SSE of the Hawkes with power-law decaying trigger density and recursive model is marginal. RMSE from overprediction and underprediction are similar for the three point process models, compared to the SEIR model which almost exclusively underpredicts as seen by the SSE % of 7.51% from overprediction.

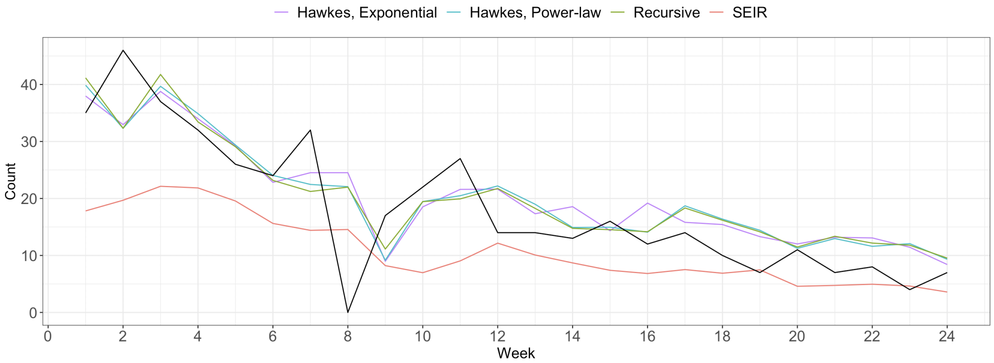


Figure 5.4: Weekly case estimates for point process and SEIR models

The error (predicted - actual) from the weekly estimates is displayed in Figure 5.5. From this plot, we can see that in general all three point processes seem to underpredict for the first 12 weeks and then overpredict in the last 12 weeks.

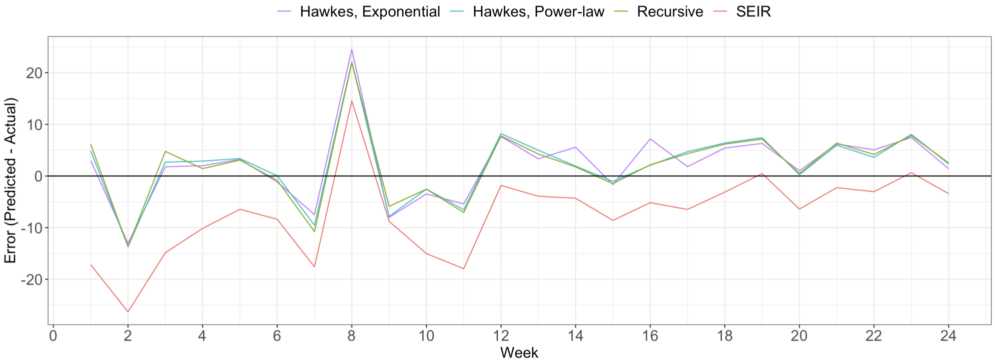
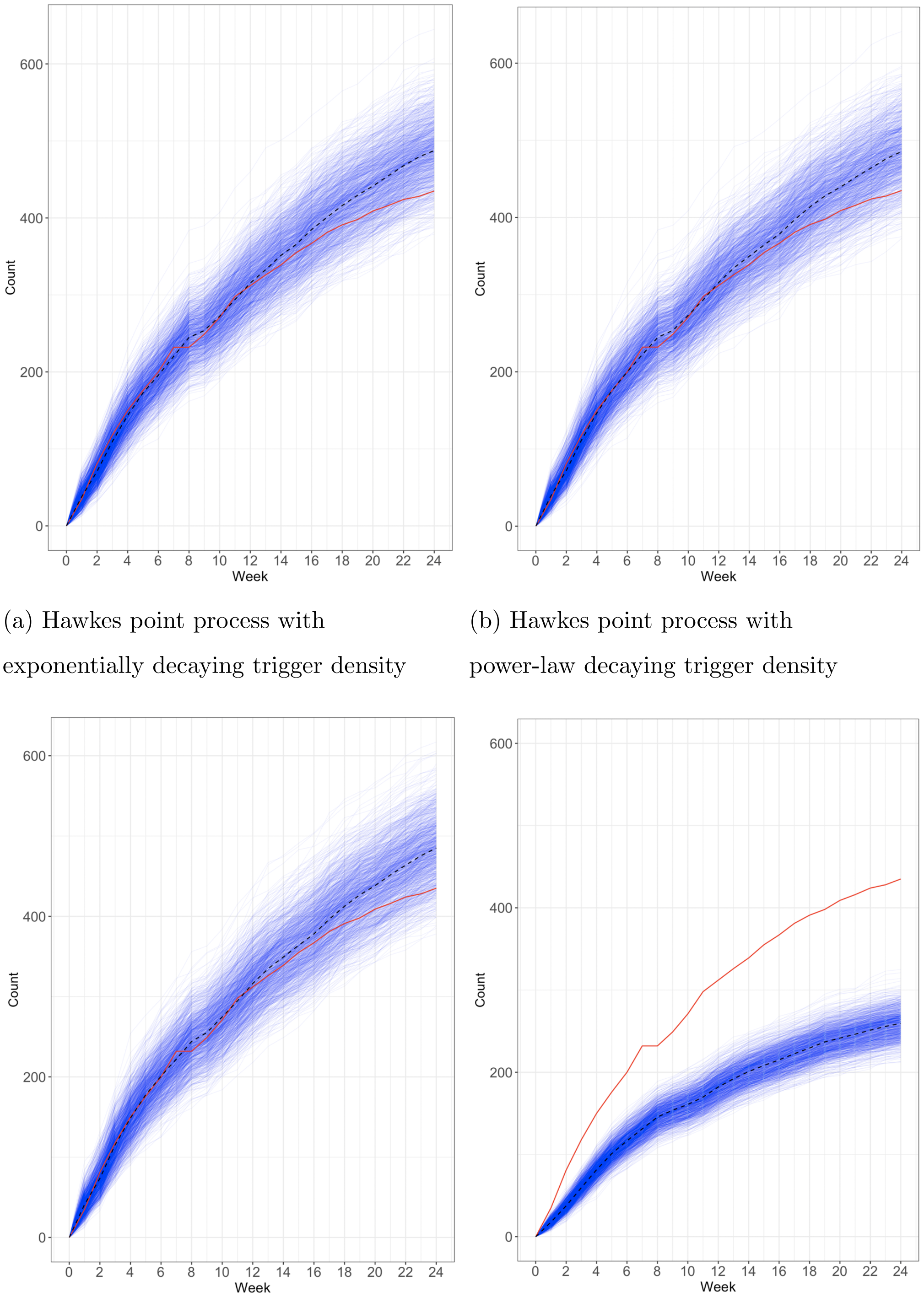


Figure 5.5: Weekly error for point process and SEIR models

Figure 5.6 shows the simulation plots for the point processes and SEIR model. The red line represents the actual cumulative weekly case count, each blue line is one of the $1,000$ simulations and the dashed black line represents the mean of the $1,000$ simulations. We can see that the variation of simulations is greater in the point processes compared to the SEIR model. Additionally, the point processes seem to match the observed cases well for the first 12 weeks, with approximately half of the simulations overestimating and half underestimating, compared to the simulations for the SEIR model which underestimate for all 24 weeks.



(a) Hawkes point process with exponentially decaying trigger density

(b) Hawkes point process with power-law decaying trigger density

(c) Recursive point process

(d) SEIR

Figure 5.6: Cumulative weekly cases of IPD

CHAPTER 6

Conclusion

From our analysis, the likelihood ratio test suggests that the additional flexibility that the recursive model offers is an improvement compared to the Hawkes with exponentially decaying trigger density. Additionally, it appears that the recursive model is able to strike a balance between fitting the training set and predicting weekly counts in the test set when compared to the other two Hawkes processes. Although the Hawkes model with power-law decaying trigger density had the lowest RMSE, it also had the highest AIC out of all the point processes. On the other hand, the recursive model has the lowest AIC and a RMSE that is 0.021 higher than the Hawkes model with power-law decaying trigger density. In comparison, the Hawkes model with exponentially decaying trigger density has a RMSE that is 0.13 higher. Therefore, in some respects, the recursive model could be viewed as a better model overall.

A limitation of our analysis was the missing weekly case counts in the pre-compiled IPD dataset from Project Tycho. The Project Tycho dataset documentation states it does not include time intervals for which no case count was reported and that the count time series are often incomplete. Consequently, for instances where there is missing data for a couple weeks, we can not be confident whether there was actually no cases during that time period or the count was simply not reported. In some cases it was more evident that the case count was simply not reported, such as the missing weekly counts between 01/01/2003 and 01/01/2004. As a result of missing data, we were only able to use approximately $4\frac{2}{3}$ out of the total 15 years in our analysis. In addition, because we only had weekly counts, we estimated occurrence times.

Subsequent analysis could be done using the point process models to forecast infections where the PCV7 and PCV13 vaccine is not readily available like certain countries in Africa or Asia. It would be informative to see how the point process performs under different circumstances. To improve the accuracy of the Hawkes, we can further investigate different trigger densities and their effect on the model performance and forecasting. One option could be the rayleigh kernel, $RAY(\gamma, \eta) = \gamma t e^{-\eta t^2}$, which has been used in the context of survival times over diffusion networks for modeling a distinct, non-monotonically decaying, type of influence (Lima & Choi, 2018). Gomez-Rodriguez, Balduzzi and Scholkopf (2011) modeled diffusion processes as discrete networks of continuous temporal processes occurring at different rates. One of the models they considered was the Rayleigh model which has been previously used in epidemiology (Rodriguez, Balduzzi, & Schölkopf, 2011). Additionally, we could compare the performance of a model developed by Rizoiu (2018) called HawkesN, which accounts for finite population sizes, with the four models in this study. The event rate function in HawkesN is defined as:

$$\lambda^H(t) = \left(1 - \frac{N_t}{N}\right)\left[\mu + \sum_{t_j < t} \phi(t - t_j)\right]$$

where $\phi(t - t_j)$ can be the same kernel function used with Hawkes, and $N(t)$ is the counting process associated with the point process (Rizoiu, Mishra, Kong, Carman, & Xie, 2018). The effect of introducing the finite population size N is that the event rate at time t is modulated by the available population (Rizoiu et al., 2018).

The SEIR model did not perform as well as the other three models. This could be a result of the chosen transmission rate $\beta(t)$ which exponentially decays. Since IPD seems to have some seasonality, a better choice for $\beta(t)$ may be a non-negative periodic function (Wang & Zhao, 2008). For example, Dietz (1976) used a SIR model with contact rate $\beta(1+\delta cos(2\pi t))$. The infectious period for IPD is presumed to last until discharges from mouth and nose no longer contain pneumococci in significant numbers. IPD can have varying infectious period depending on the serotype; this makes it hard to determine an accurate infection rate γ for our population, the state of Florida (Sleeman et al., 2006). Additionally there is evidence that certain serotypes are more common in children younger than 16 years old compared to adults 16 years or older (Imöhl, Reinert, Ocklenburg, & van der Linden, 2010). The assumptions we made in the SEIR model may not have been the most appropriate for describing the dynamics of IPD. Melegaro et al. (2010) developed, parameterized, and applied an age-structure transmission dynamic model that also included heterogeneous-mixing and predicted the overall incident of IPD in England and Wales. Further investigation of different models and their ability to accurately forecast the spread of IPD is needed to provide better insight into IPD dynamics and improve forecasting performance which in turn can guide vaccination recommendations.

References

Advisory Committee on Immunization Practices. (2000). Preventing pneumococcal disease among infants and young children. recommendations of the advisory committee on immunization practices (acip). *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports*, *49*(RR-9), 1.

Althaus, C. L. (2014). Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. *PLoS currents*, *6*.

Anderson, R. M. (1991). Discussion: the kermack-mckendrick epidemic threshold theorem. *Bulletin of mathematical biology*, *53*(1-2), 1.

Bacry, E., Mastromatteo, I., & Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, *1*(01), 1550005.

Baddeley, A., Turner, R., Møller, J., & Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(5), 617–666.

Bravo, L. (2009). Overview of the disease burden of invasive pneumococcal disease in asia. *Vaccine*, *27*(52), 7282–7291.

Bridy-Pappas, A. E., Margolis, M. B., Center, K. J., & Isaacman, D. J. (2005). Streptococcus pneumoniae: description of the pathogen, disease epidemiology, treatment, and prevention. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *25*(9), 1193–1212.

Bureau of Economic and Business Research, & University of Florida. (2014). *Florida estimates of population 2014*. University of Florida. Retrieved from `https://www.bebr.ufl.edu/sites/default/files/Research%20Reports/estimates_2014.pdf`

Clements, R. A., Schoenberg, F. P., & Veen, A. (2012). Evaluation of space–time point process models using super-thinning. *Environmetrics*, *23*(7), 606–616.

Daley, D. J., & Vere-Jones, D. (2003). An introduction to the theory of point processes, volume 1: Elementary theory and methods. *Verlag New York Berlin Heidelberg: Springer*.

Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T., & Jacobsen, K. H. (2019).

Complexity of the basic reproduction number (r0). *Emerging infectious diseases*, *25*(1), 1.

Everett, D. B., & et al. . (2011). Ten years of surveillance for invasive streptococcus pneumoniae during the era of antiretroviral scale-up and cotrimoxazole prophylaxis in malawi. *PloS one*, *6*(3), e17765.

Gumel, A. B., Ruan, S., Day, T., Watmough, J., Brauer, F., Van den Driessche, P., . . . Sahai, B. M. (2004). Modelling strategies for controlling sars outbreaks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1554), 2223–2232.

Hamborsky, J., & Kroger, A. (2015). *Epidemiology and prevention of vaccine-preventable diseases, e-book: The pink book*. Public Health Foundation.

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

Imöhl, M., Reinert, R. R., Ocklenburg, C., & van der Linden, M. (2010). Association of serotypes of streptococcus pneumoniae with age in invasive pneumococcal disease. *Journal of clinical microbiology*, *48*(4), 1291–1296.

Johnson, P. (2014). adaptivetau: efficient stochastic simulations in r. *R Package Version*, 2–2.

Laub, P. J., Taimre, T., & Pollett, P. K. (2015). Hawkes processes. *arXiv preprint arXiv:1507.02822*.

Li, M. Y. (2018). *An introduction to mathematical modeling of infectious diseases* (Vol. 2). Springer.

Lima, R., & Choi, J. (2018). Hawkes process kernel structure parametric search with renormalization factors. *arXiv preprint arXiv:1805.09570*.

Melegaro, A., Choi, Y. H., George, R., Edmunds, W. J., Miller, E., & Gay, N. J. (2010). Dynamic models of pneumococcal carriage and the impact of the heptavalent pneumococcal conjugate vaccine on invasive pneumococcal disease. *BMC infectious diseases*, *10*(1), 90.

Mohler, G. O., 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.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, *7*(4), 308–313.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, *50*(2), 379–402.

Organization, W. H. (2016). Pneumococcal meningitis outbreaks in sub-saharan africa [Journal / periodical articles]. *Weekly Epidemiological Record*, *91*(23), 298 - 303.

Ozaki, T. (1979). Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, *31*(1), 145–155.

Park, J., Chaffee, A. W., Harrigan, R. J., & Schoenberg, F. P. (2020). A non-parametric hawkes model of the spread of ebola in west africa. *Journal of Applied Statistics*, 1–17.

Pneumococcal disease. (2017). National Center for Immunization and Respiratory Diseases, Division of Bacterial Diseases. Retrieved from `https://www.cdc.gov/pneumococcal/surveillance.html`

Rivers, C. M., Lofgren, E. T., Marathe, M., Eubank, S., & Lewis, B. L. (2014). Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia. *PLoS currents*, *6*.

Rizoiu, M.-A., Mishra, S., Kong, Q., Carman, M., & Xie, L. (2018). Sir-hawkes. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*.

Rodriguez, M. G., Balduzzi, D., & Schölkopf, B. (2011). Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697*.

Schoenberg, F. P. (2010). Introduction to point processes. *Wiley Encyclopedia of Operations Research and Management Science*.

Schoenberg, F. P., Hoffmann, M., & Harrigan, R. J. (2019). A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics*, *71*(5), 1271–1287.

Sleeman, K. L., Griffiths, D., Shackley, F., Diggle, L., Gupta, S., Maiden, M. C., … Peto, T. E. (2006). Capsular serotype–specific attack rates and duration of carriage of streptococcus pneumoniae in a population of children. *The Journal of infectious diseases*, *194*(5), 682–688.

van Panhuis, W. G., Cross, A., & Burke, D. S. (2018). Project tycho 2.0: a repository to improve the integration and reuse of data for global population health. *Journal of the American Medical Informatics Association*, *25*(12), 1608–1617.

Wang, W., & Zhao, X.-Q. (2008). Threshold dynamics for compartmental epidemic models in periodic environments. *Journal of Dynamics and Differential Equations*, *20*(3), 699–717.

Weinbergera, M., Malley, R., & Lipsitch, M. (2012). Serotype replacement in disease following pneumococcal vaccination: A discussion of the evidence. *Lancet*, *378*, 1962–1973.