Title
Comparative Analysis of SEIR and Hawkes Models for the 2014 West Africa Ebola Outbreak

Author
Chaffee, Adam

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Comparative Analysis of SEIR and Hawkes Models for the 2014 West Africa Ebola Outbreak

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Adam Walter Chaffee

2017

ABSTRACT OF THE THESIS

Comparative Analysis of SEIR and Hawkes Models for the 2014 West Africa Ebola Outbreak

by

Adam Walter Chaffee

Master of Science in Statistics

University of California, Los Angeles, 2017

Professor Frederic R. Paik Schoenberg, Chair

The extent to which Hawkes point process models can more accurately characterize the evolution of a disease epidemic than a standard compartmental model such as SEIR is investigated. Maximum likelihood estimation was used to fit SEIR model parameters to Ebola outbreak data in West Africa in 2014 from the World Health Organization (WHO). Projections using simulation were then conducted using the Poisson-leaping Tau Method (Cao et al. 2007) to evaluate the fit. The projections and rate function were compared to Hawkes point process estimation and simulation over the same data and projection scale. Results indicate that Hawkes models outperformed SEIR in predicting the spread of Ebola in West Africa with a 38% reduction in RMSE for weekly case estimation across all countries (total RMSE of 59.6 cases/week using SEIR compared to 37.2 for Hawkes). An analysis using the first 75% of the data for estimation and the subsequent 25% of the data for evaluation shows that the improved fit from Hawkes modeling cannot be attributed to overfitting.

The thesis of Adam Walter Chaffee is approved.

Nicolas Christou

Ying Nian Wu

Robert Gould

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2017

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I am extremely grateful for the contributions from my advisor and committee chair, Professor Rick Paik Schoenberg. His teaching at UCLA provided the initial spark of interest in spatial probability models, and the help and guidance he provided for this thesis was invaluable. I would also like to thank my committee members: Nicolas Christou, Robert Gould, and Ying Nian Wu. Thank you for being such diligent reviewers to make this paper the best it can be.

I would also like to acknowledge the contributions from Ryan Harrigan, Alex Krebs, and Junhyung Park, who along with Rick provided guidance for this work and are utilizing the results in preparation for an academic publication. I would especially like to thank and formally acknowledge Junhyung as a significant contributor to this paper for providing Hawkes analysis that produced Hawkes model numbers in Table 1, as well as Hawkes results in Figures 1, 2, 5, and 7.

Finally, and most importantly, I would like to thank my best friend and life partner, Lauren. Her support, encouragement, and love mean the world to me and motivated me to persevere through the rigors of graduate school. I also am grateful to my parents, Doug and Paulette, for going the extra mile (literally and figuratively) in supporting me to do my best and pursue my dreams since childhood.

Introduction

Between March 2014 and June 2016, the West African countries of Guinea, Sierra Leone, and Liberia fell victim to an Ebola outbreak of massive scale that rapidly grew to become larger than all other previously recorded Ebola outbreaks combined (WHO Ebola Response Team, 2014). With nearly 30,000 total suspected Ebola cases and more than 11,000 deaths (World Health Organization, 2016), the outbreak severely diminished each country's quality of life and economic output due to decreased trade, border closures, and a drop in foreign investment (United Nations Development Programme, 2015). The West African Ebola outbreak serves as a stark reminder that major outbreaks of deadly diseases can still occur with major impacts to life and socioeconomic well-being.

To prevent future outbreaks of Ebola and other highly infectious diseases, it is important that communities and governments focus on several areas including improving detection and response capacity, conducting survivor studies, providing adequate education to the public, and continuing support for disease research (Spengler et al., 2016). Statistics can provide great contributions to many of these areas.

In particular, statistical models can help predict the spread of infectious diseases once an outbreak begins, leading to more effective allocation of detection and response resources. One of the first major breakthroughs in epidemiological modeling was the development of the compartmental model by Kermack and McKendrick (1927) which led to the basic SIR (Susceptible-Infected-Recovered) model and its variants. Such models involve dividing populations according to disease status, and then modeling the changes in numbers of infected, susceptible, and recovered individuals in the population using systems of simple differential equation models. Compartmental modeling has grown to become a mainstay of the

epidemiological community for modeling the establishment and spread of infections. In recent decades, the traditional SIR model has been modified with differing levels of complexity to better fit individual disease characteristics (Britton, 2010).

The SEIR (Susceptible-Exposed-Infected-Recovered) compartmental model has become especially popular for describing the dynamics of the Ebola virus, most notably by Chowell et al. (2004) and applied to West African Ebola data by Althaus (2014). However, SEIR and other compartmental models may be overly simplistic in that they rely on spatial aggregation and subsequently describe the purely temporal spread, rather than describing the spatial-temporal interaction in detail. In addition, such models rely on the mass action assumption (Meyers, 2007) that all susceptible members of the population are equally likely to be infected, which is typically violated in practice, often resulting in inaccurate forecasts. For instance, compartmental SEIR models applied to the spread of SARS in China in 2003 estimated a high transmission rate and suggested 30,000 to 10 million SARS cases would occur in the first 4 months of the spread of disease in China, resulting in fears of a widespread pandemic (Meyers, 2007). Ultimately, only about 5,300 cases were reported in China (World Health Organization, 2003).

Much work has been done in recent years pertaining to parameter estimation and theoretical modeling for Ebola SEIR models, yet comparatively little attention has been paid to assessing the fit of SEIR models to real data and comparing them with alternatives. This paper attempts to address this discrepancy by applying SEIR modeling techniques to the beginning of the 2014 West Africa Ebola outbreak, then utilizing these models to evaluate the goodness of fit on the data itself through case count estimation via mean simulation results. The results are then compared to goodness of fit of an alternative technique, Hawkes point process modeling, using equivalent data.

Hawkes models are particularly suitable for describing the processes by which humans spread contagious diseases. Such processes were used to model the occurrence of smallpox in Brazil by Becker (1977), and by Farrington et al. (2003) to describe the effect of vaccinations on cases of measles in the United States, but their use for describing the emergence and spread of infectious diseases so far remained scant. As an alternative to compartmental SEIR models and their variants, Hawkes models may provide new insights into the spread of epidemics and invasive species, including a description of the spread via an estimated triggering kernel.

An overview of Hawkes model fitting is presented and compared to SEIR model fitting developed in this paper. Hawkes models are often extended to spatial data in space and time such as infection times and locations. However, to provide an equivalent comparison to SEIR modeling which is assumed spatially homogeneous, the Hawkes model was fit to infection times only. The fitted results of each modeling technique are then compared to assess their relative advantages and shortcomings when applied to various time intervals during the early stages of the 2014 West Africa Ebola outbreak. This work is a case study to provide a preliminary assessment of how SEIR compartmental models and Hawkes point process models vary in effectiveness in predicting the magnitude of infections in disease outbreaks.

Design

Outbreak Data

Data was collected and aggregated from the World Health Organization (WHO) website (http://www.who.int/en/) outbreak reports on Ebola during and after the outbreak period. These reports were typically released weekly by WHO and included the country, geographic location within country (either by region, closest city, or village) as well as confirmed cases and deaths from Ebola virus. For all locations, geographic coordinates were determined by searching for

village names either in Google Earth (Google Earth 2015, https://www.google.com/earth/download/ge/) or using the FallingRain Gazeteer (http://www.fallingrain.com/world/index.html) and recording the latitude and longitude of the location. In cases where only region was reported, we identified the centroid of that region and used the latitude and longitude of that centroid as the location for all cases.

Data was filtered to only include infection cases and death counts from Ebola at various measured time points in three regions: Southeast Guinea, Eastern Sierra Leone, and Northwest Liberia. The time range of these observations begins on March 23, 2014 and ends on September 7, 2014. This time window was used because it is similar to the data used from SEIR model fitting conducted in Althaus (2014), and will facilitate equivalent comparisons using Hawkes modeling. A copy of the data used is produced in the Appendix.

The filtered data was modified to approximate real-time person-by-person infections by uniformly distributing new infections within times between reporting dates. Cases at the initial reporting date were assumed to have occurred uniformly over 1 day prior to the initial date. This modified data was used in fitting the Hawkes model parameters and used in all evaluation techniques for both Hawkes and SEIR models. The original filtered data was only used in SEIR model fitting.

SEIR Modeling Overview

The SEIR (Susceptible-Exposed-Infected-Recovered) compartmental model embodies the idea that the infected population spreads the disease at time t with rate $\beta(t)$, but can only spread the disease to the proportion of the population still susceptible, and these rates and proportions can change as an outbreak proceeds. It has been frequently used to describe Ebola

disease dynamics and is characterized by the following set of ordinary differential equations (Chowell et al. 2004):

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

Where *S* is the susceptible population, *E* is the population which has contracted Ebola but is not yet infectious ("latent population"), *I* is the infectious population, and *R* is the recovered/deceased population. These four quantities sum to *N*, the total population. The populations are assumed to be spatially homogeneous, with a fixed *N* over time.

When modeling the infectious phase, the primary quantity of interest in this model is $\beta(t)$, the transmission rate. Under this model, it is assumed to decline exponentially at rate κ:

$$\beta(t) = \beta e^{-\kappa t}$$

Where *t* is the number of days from the start of the outbreak (Lekone and Finkenstädt, 2006). Other parameters in the SEIR model include the rate of infectious onset, σ, and the rate of death or recovery, γ. In model fitting, σ and γ are assumed constant to replicate the techniques used by Althaus (2014).

A central feature to compartmental SIR/SEIR modeling is the reproductive number, R_0. In the model R_0 at any given time is estimated by the transmission rate, $\beta(t)$, multiplied by the average duration of infectiousness, 1/γ. R_0 represents the average number of new infections generated by an infected person until the infected person dies or recovers. The critical threshold for R_0 is 1: if R_0 is above 1, the epidemic can spread to infect a large proportion of the population. When R_0 drops below 1, the epidemic will quickly stop.

The SEIR model was fitted separately to Guinea, Sierra Leone, and Liberia using the original discretely reported Ebola outbreak data containing only cases and deaths at each reporting date. This approach is identical to the SEIR model used by Althaus (2014) in his parameter estimation for the outbreaks in Guinea, Sierra Leone, and Liberia.

It is important to note that the theoretical SEIR model outlined above is purely deterministic. To convert this process into a stochastic model for simulating real-world outbreaks, the Tau-leaping approximation (Cao et al., 2007) is applied to the model. Under this process new exposures, infections, and recoveries occur randomly as a Poisson arrival process at probabilities based on R_0, κ, σ, and γ. With each new transition, the probabilities of these events update to reflect the new *S, E, I,* and *R* populations. The stochastic model using Tau-leaping is used in weekly and overfitting estimates as outlined in the Evaluation Techniques.

Hawkes Modeling Overview

A point process (Daley and Vere-Jones, 2003, 2007) is a collection of points $\{\tau 1, \tau 2, ...\}$ occurring in some metric time space. The points can be defined in space, time, or both, but for this study the points are modeled in time only. Such processes are typically modeled via their conditional rate (also called conditional intensity), $\lambda(t)$, which represents the infinitesimal rate at which points are accumulating at time t, given information on all points occurring prior to time t.

Hawkes or self-exciting point processes (Hawkes, 1971) are a type of branching point process model that has become widely used in modeling seismicity (Ogata, 1988, 1998) and other natural phenomena, but have scarcely been used to describe the spread of epidemic diseases. Hawkes models are frequently applied to data in both space and time. In this study, however, Hawkes models exclude spatial components to provide an equivalent comparison to the

spatially homogeneous SEIR model. For a purely temporal Hawkes process, the conditional rate of events at time t, given information H_t on all events prior to time t, can be written:

$$\lambda(t|H_t) = \mu(t) + \kappa \sum_{\{t':\ t'<t\}} g(t-t')$$

Where $\mu(t)$ is the background rate, or the rate of infections that develop due to random chance without transmission from another infected individual. The quantity g is the triggering function, and κ is the productivity. The triggering function can be thought of as a cumulative density function over time for the model, or the probability that an infection at time t' will trigger another infection at or before time t, where $t' < t$. As a cumulative density function, g approaches 1 with increasing time. The quantity κ is the expected number of new infections created by each individual infection. These new infections can then cause an average of κ additional infections in a cascading series. Because κ is constrained to be between 0 and 1, each background point is expected to generate a total of $1/(1-\kappa) - 1$ points by convergence of a geometric series.

Evaluation Techniques

The parameters for the SEIR model were obtained using maximum likelihood estimation (MLE) based on Althaus (2014) which assumes occurrence of new cases follow a Poisson distribution. The negative log-likelihood score is calculated by a Poisson density function comparing probabilities of observed and expected case counts. The optimization algorithm uses the techniques developed by Nelder and Mead (1965) to find model parameters which minimize negative log-likelihood values. Hawkes model parameters and likelihood scores were estimated using non-parametric maximum likelihood estimation (Marsan and Lengliné 2008). The method approximates maximum likelihood by using the E-M algorithm. Parameters are first initialized; g

is initialized as a density step function with a fixed number of steps. Given those parameters, for all pairs of points *i* and *j*, where $\tau_i > \tau_j$, the probability event *i* triggered event *j* is calculated. Each pair of points is then re-weighted by its probability. This two-step process repeats until parameters converge to a local optimum.

The SEIR fitted model parameters were then used in simulation-based forecasting for prospective analysis. Simulations were conducted in the R program for statistical computing using compartment estimation with the deSolve package, along with the tau-leaping method (Cao et al., 2007) which uses the adaptiveTau package. To begin SEIR model simulation, each population compartment must be initialized. Constant durations of infectiousness and latency of 6 days each was used to estimate the starting infectious and latent populations for simulations. The starting susceptible populations were set to equal regional populations using the most recently published census data from Guinea (National Institute of Statistics, 2015), Sierra Leone (Sierra Leone Statistics, 2016), and Liberia (LISGIS, 2009).

In the weekly projection analysis, SEIR and Hawkes MLE parameter estimates using the entire country infection data sets were used to project cumulative infections on a weekly basis using the mean of 1,000 simulations per week per country. The projection of the first week of the outbreak requires starting populations at day 0 of the outbreak, which are unknown. Therefore, during the first week's projections, infectious and susceptible population numbers were set to the observed values, and only subsequent weeks were used for evaluation.

In a separate analysis to guard against possible overfitting, parameter estimates were obtained withholding the last 25% of data, and were then used to project cumulative infections for the held-out period. To explore additional SEIR model variability, an overfitting analysis was

conducted withholding the last 50% of data. The overfitting analysis also involved 1,000 simulations per country.

In point process analysis, an effective way to evaluate the fit of the modeled rate of new events is to conduct a super-thinning analysis which assesses the homogeneity of the residuals of model rate parameters over space and/or time (Clements et al., 2013). Super-thinning involves both thinning the existing data points and superposing a new set of points. This technique was implemented in assessing goodness of fit for both Hawkes and SEIR models.

For super-thinning the Ebola outbreak in time, the estimated rate of new infections ($\widehat{\lambda(t)}$) is calculated for each infection time in the data. Superthinning requires the choice of a tuning parameter, b, and as suggested in Clements et al. (2013) we use the simple default value of the total number of cases divided by the length, in days, of the observation period. First the existing data points are thinned where each point is randomly kept with probability $\min\{b/\widehat{\lambda(t)}, 1\}$ leaving a residual process. New points are then superposed via a two-step process. A Poisson process with constant rate b is first generated over the time interval, then each point is independently kept with probability $\max\{b - \widehat{\lambda(t)}/b, 0\}$.

In evaluating the SEIR models, the value of $\beta(t)$ multiplied by the infectious population at time t was used as the estimated rate function ($\widehat{\lambda(t)}$) to calculate thinning and superposing probabilities. Hawkes models were evaluated using super-thinning based on the estimated rate function of the model ($\widehat{\lambda(t)}$).

Super-thinned and superposed points are overlaid and visualized with each time point on the x-axis. A random uniform variable was generated as a y-coordinate for each point on the plot. Different point styles were used to differentiate between thinned original infection times and superposed times. The resulting superposed process is then examined for uniformity, as the

residuals should form a stationary Poisson process if the modeled rate is correct (Clements et al. 2013). Time regions in the plot that become sparse indicate periods of overprediction because the few observed points in the region are removed with high probability, then new points are superposed with a probability close to 0. Highly clustered regions indicate underprediction.

Results

Model Fitting and Weekly Estimates

The log-likelihood scores for all models are shown in Table 1. Because Hawkes and SEIR models rely on different underlying probabilistic processes, the individual likelihood of a given model is difficult to interpret substantively, but the fit of two distinct models can easily be compared using log-likelihoods. In nested point process models, for example, the difference in log-likelihood of nested models is approximately chi-square distributed with 2q degrees of freedom, where q is the difference in the number of parameters between the two models (Ogata, 1978). Here the models are not nested, so a common alternative in maximum likelihood estimation is to calculate the AIC when comparing models. AIC is calculated as {-2*log-likelihood + 2*p} where p is the number of estimated parameters in the model. Lower AIC indicates better fit.

The AIC for all three countries is lower in the Hawkes model, indicating that the Hawkes models provided a better fit to the infection outbreak in all three countries. When comparing root mean square error (RMSE) of weekly predicted cases in Table 1, Hawkes models achieved lower RMSE than SEIR models for all countries. The total RMSE across all countries was 59.6 cases/week using SEIR and 37.2 cases/week using Hawkes models which represents a 38% decrease. RMSE and sum-squared error (SSE) attributable to overprediction and underprediction are shown in Table 1 as well. SEIR had greater RMSE in all countries and categories except in

underpredictions for Liberia. The RMSE results in total listed above and in Table 1 indicate that Hawkes achieved better model fitting than SEIR for nearly all data in the study.

Table 1. Log-likelihood, AIC, and squared error for SEIR and Hawkes model fitting

	Guinea (861 cases)		**Sierra Leone (1424 cases)**		**Liberia (2081 cases)**	
	SEIR	Hawkes	SEIR	Hawkes	SEIR	Hawkes
Log-Likelihood	-606.5	913.6	-239.3	2834.8	-330.0	5265.4
AIC	-1207.0	-1223.2	-472.6	-5065.6	-654.0	-9926.8
Weekly Prediction Results						
RMSE from prediction	33.1	17.4	92.2	51.8	54.3	41.2
SSE% from over-predicting	32.9%	48.9%	55.2%	18.8%	64.2%	52.6%
SSE% from under-predicting	67.1%	51.1%	44.8%	81.2%	35.8%	47.4%
RMSE from over-predicting	23.0	15.1	97.0	31.2	74.7	35.3
RMSE from under-predicting	47.0	21.1	78.8	66.3	41.6	53.0

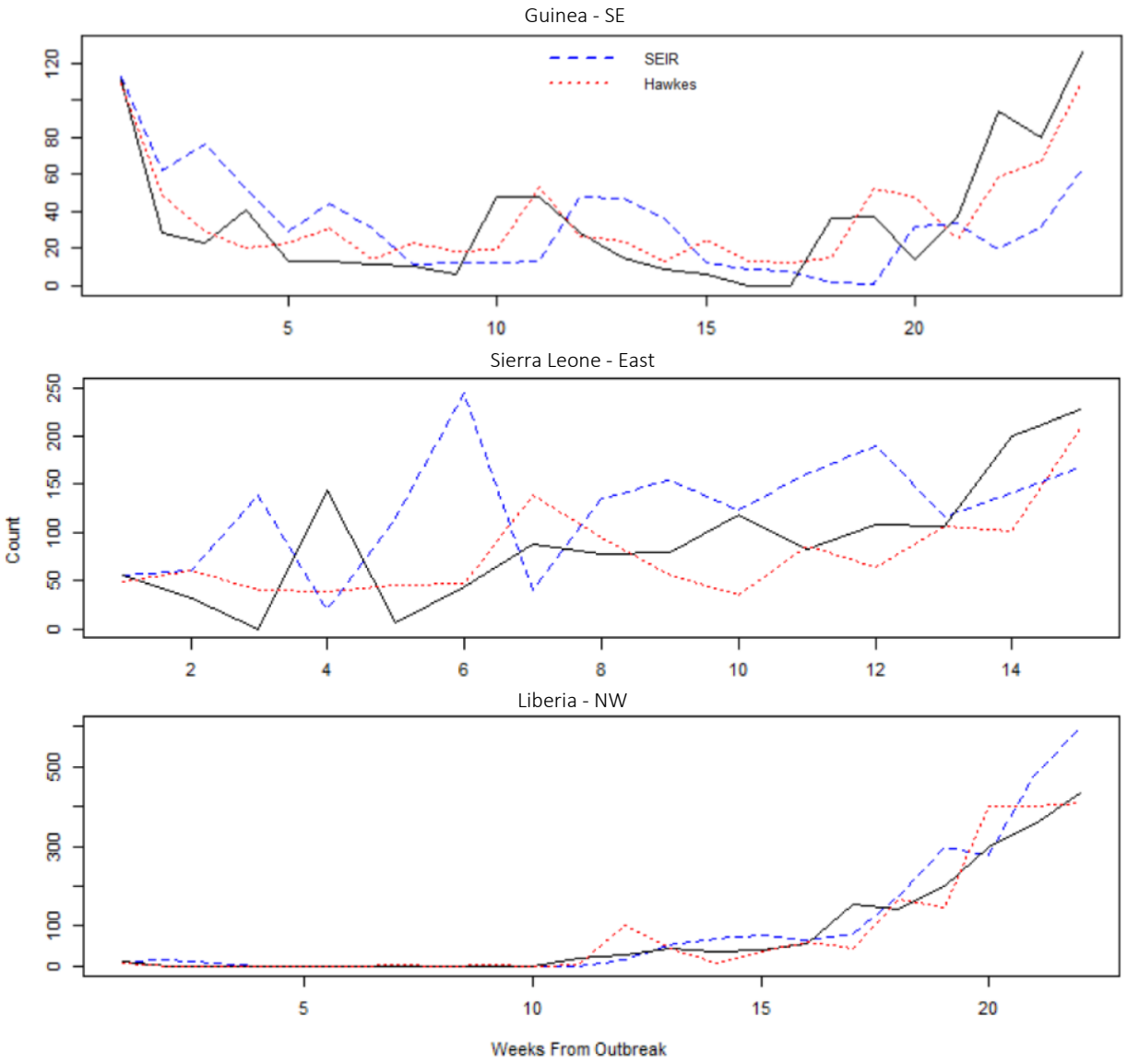
The results from weekly estimates of total infections per week are displayed in Figure 1. Hawkes modeling projections for each week is shown in blue, SEIR is in red, and the actual data is shown in black. The projections from SEIR models display a strong relatedness to case load from prior weeks. This is not surprising given that the rate of infectiousness for SEIR modeling is a function of the current infectious population. SEIR model estimates tend to lag by nearly two weeks due to the assumed incubation period of 6 days before reaching infectiousness, which then lasts an additional 6 days.

A general feature of the SEIR weekly estimates in Guinea and Sierra Leone is the tendency for the model to over predict in earlier weeks, then under predict in later weeks. This is caused by the nonzero κ term causing an inverse exponential decrease in the assumed transmission rate, $\beta(t)$. In Liberia, the maximum likelihood estimate for κ is 0, so there tends to be a consistent level of overprediction in case load except when there are sudden, large jumps in case counts from week to week.

Hawkes models show a similar dependence due to the productivity constant in the model. However, the dependence is much weaker possibly because the background rate of the model is

calculated over the entire course of the data. This may provide a benefit of excluding some noise produced from week-to-week infection variability. Across all three countries the Hawkes models tend to produce more accurate weekly estimates with less extreme errors.

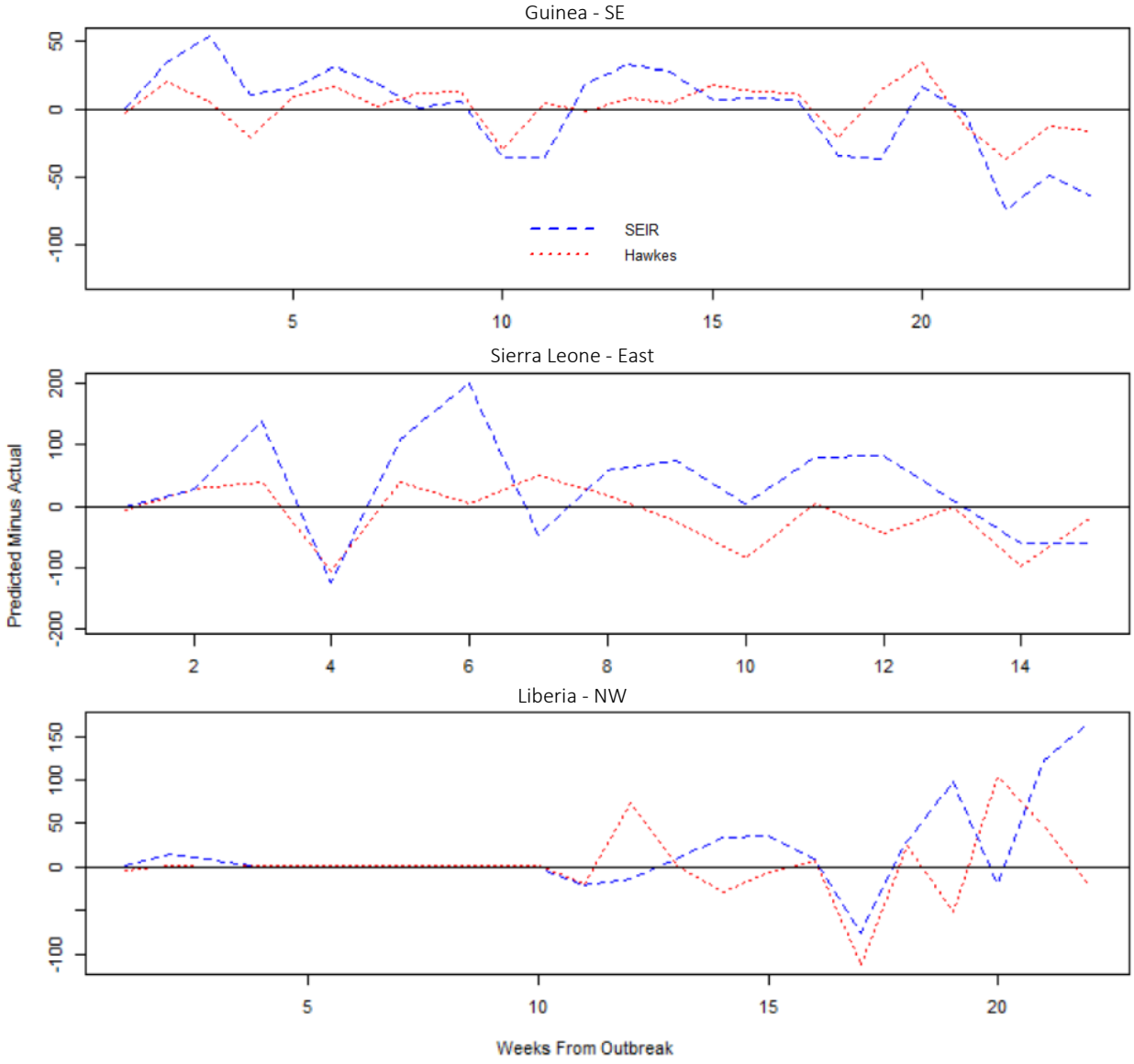
Figure 1. Weekly case estimates from SEIR and Hawkes models for all countries



The error (estimated cases – actual cases) from weekly estimates is displayed for all models in Figure 2. Hawkes model error is shown in blue, SEIR is shown in red, and the black

horizontal line at 0 represents perfect estimates. The error rates for SEIR tend to be slightly more variable, but overall the prediction error is on the same scale as Hawkes. The analysis of weekly estimates indicate that Hawkes models tend to perform better at estimating caseloads one week into the future.

Figure 2. Weekly error from SEIR and Hawkes models for all countries



Prospective Analysis

Simulation plots for fitted SEIR parameters on the first 50% and 75% of outbreak time data to the remaining 50% and 25%, respectively, are shown in Figures 3 and 4. The thick red

line indicates the actual cumulative occurrence of cases over time. Each of the thin transparent blue lines represent one of the 1,000 simulations of caseload over time which collectively provide a sense of how well or poorly each model explained the held-out data. A red dashed line is also overlaid which represents the mean of the 1,000 simulations.

In the 50% fitting and projecting, the case load for Guinea was initially overpredicted, then a sharp increase in actual case load was not expected by any of the simulations. This led to underestimation for most of the simulations towards the end of the time. Sierra Leone simulations showed an average slight overestimation throughout the course of the time, but the data generally lies within the simulation variability. Liberia's caseload is consistently underestimated in simulations throughout the held-out period.

In the 75% fitting and projecting, SEIR modeling significantly underestimated Guinea's new cases throughout the course of the simulation. The SEIR model for Sierra Leone also underestimated, but produced simulations that were only slightly lower than the actual trajectory of new cases. Sierra Leone's simulations also captured actual results quite well during the first 2 weeks of time. Liberia's SEIR model simulations first undershot the actual case load after the first week but then began overestimating from exponential acceleration in all simulations. This acceleration did not match well with the observed linear increase.

The simulations from all three countries achieved a varying degree of success in capturing actual case incidence. The amount of data used to fit SEIR models had a large impact in determining the direction and variability of simulations. In addition, simulations were impacted by the estimated populations of infected individuals for each country. If a model was fit at a time point right after a reported sharp increase in case load, such as the Liberia 75% fit, the simulations would project large increases. Given a time period culminating in few new

infections, the SEIR model tended to forecast a continued period of very few new infections, which was inconsistent with the actual observations.

Figure 3. SEIR projections using 50% of data

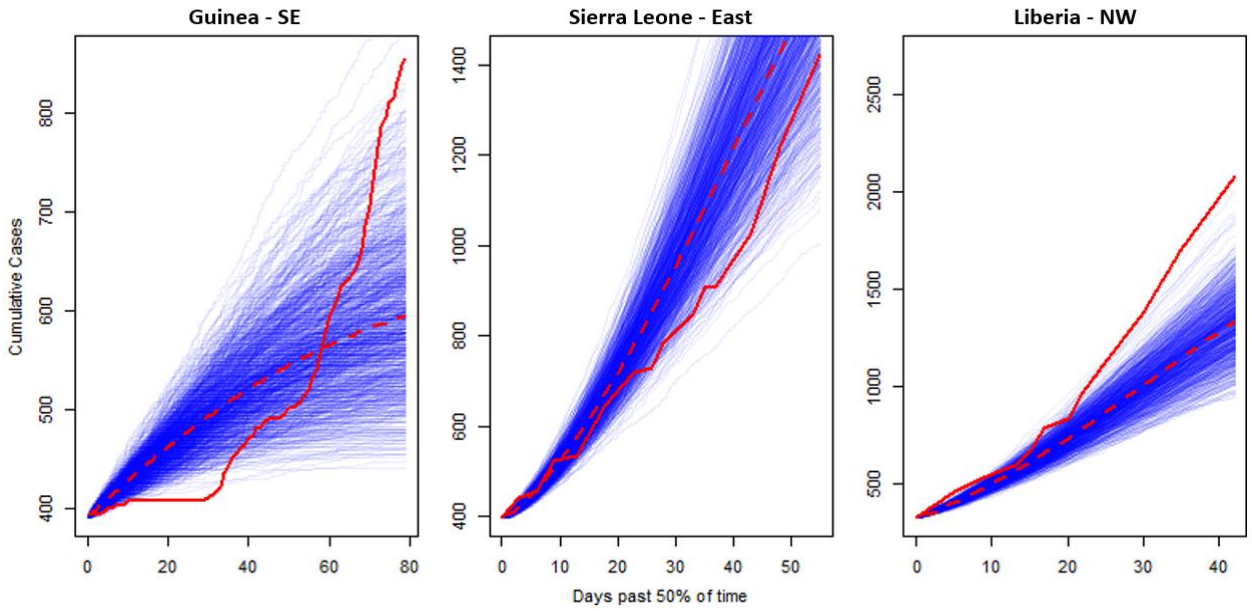


Figure 4. SEIR projections using 75% of data

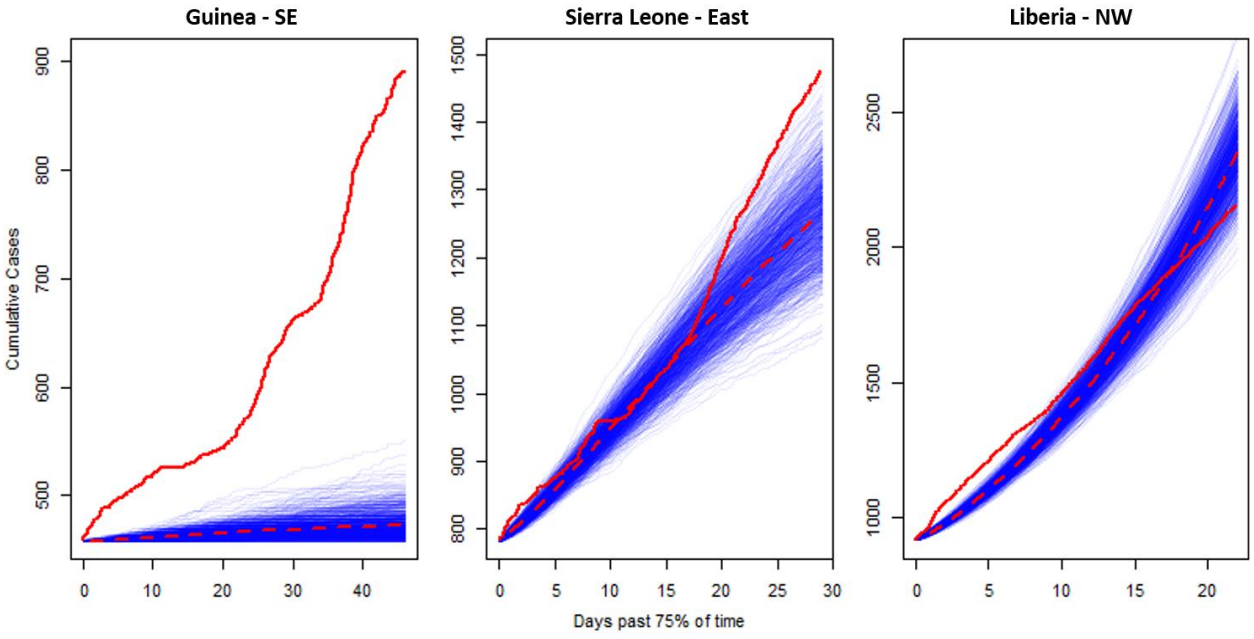
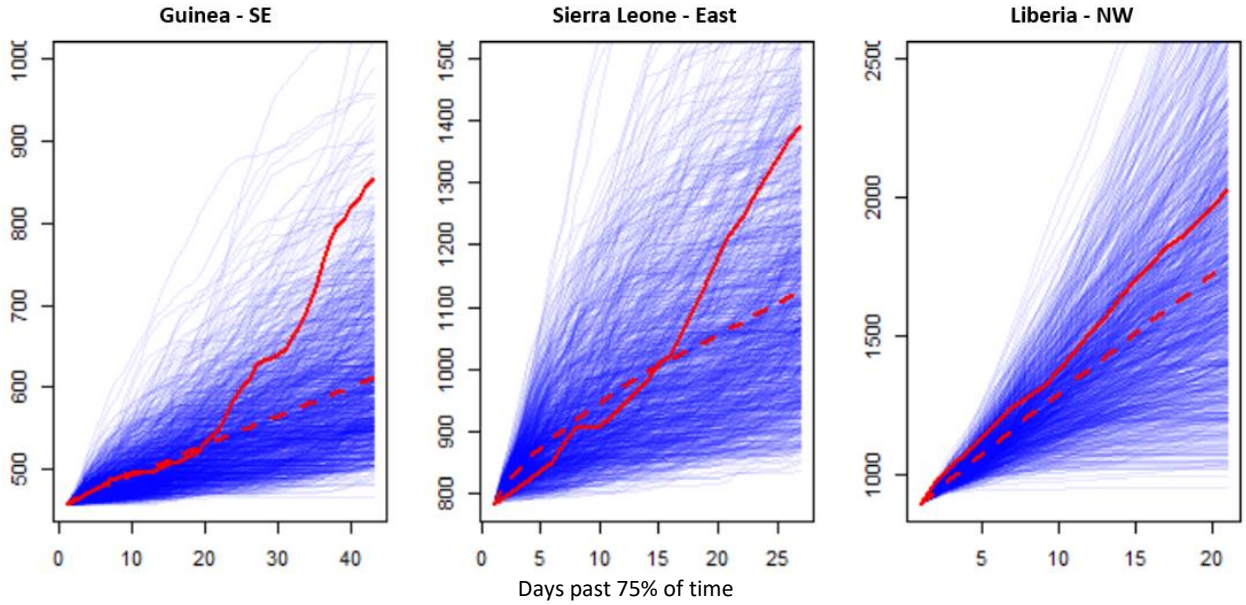


Figure 5 shows equivalent 75% fit and simulated projections using the Hawkes model. Hawkes simulations mostly underpredicted Guinea's cases, although a small proportion

generated slight overestimation. Sierra Leone's Hawkes estimates in aggregate seem to match the actual data well with about half of simulations overestimating caseloads and half underestimating, but the variation of simulations is much greater compared to the SEIR model. Liberia's simulations also exhibit large variation but in aggregate fit the held-out data reasonably well. For these data and simulations, Hawkes seemed to overall provide a better fit to the observed data for each country, although the variance in simulation results was generally much larger than for the SEIR models.

Figure 5. Hawkes projections using 75% of data



Super-thinning Analysis

Super-thinning results are displayed in Figures 6 and 7 for all regions. Original thinned points are displayed as "+" symbols and superposed points are represented by "o" symbols. The SEIR super-thinning plots in Figure 6 lack homogeneity in certain time ranges.

In SEIR super-thinning, Guinea shows high clustering and therefore extreme underestimation during the first week, as well as around Day 70 and Day 150. Overestimation from a lack of points is seen in several places in Guinea but most noticeable around Day 95. Sierra Leone's super-thinning indicate poor fit in many places as well but most notably

underprediction near Day 90, and overprediction near Day 100. Liberia's super-thinning plot appears like a Poisson process with constant rate throughout the course of the outbreak data.

The super-thinned residuals for the SEIR model clearly indicate excessive clustering. This likely occurs because the rate function is heavily dependent on the current infectious population which can be highly variable. When there is an unexpected surge in observed infections, the modeled rate according to the SEIR model tends to remain relatively low for several weeks. As a result, most of the observed points are retained after superthinning, resulting in intense clustering. After several weeks, the SEIR rate increases rapidly, resulting in very few points retained following superthinning.

The Hawkes super-thinning plots in Figure 7 appear to be consistent with Poisson processes with constant rate for all three countries. We observe no evidence that the model is inaccurately estimating the infection rates over time. The superthinning indicates that Hawkes models tended to describe the rate of new infections throughout the course of the outbreak more accurately.

Figure 6. Super-thinning using SEIR infection rate parameter

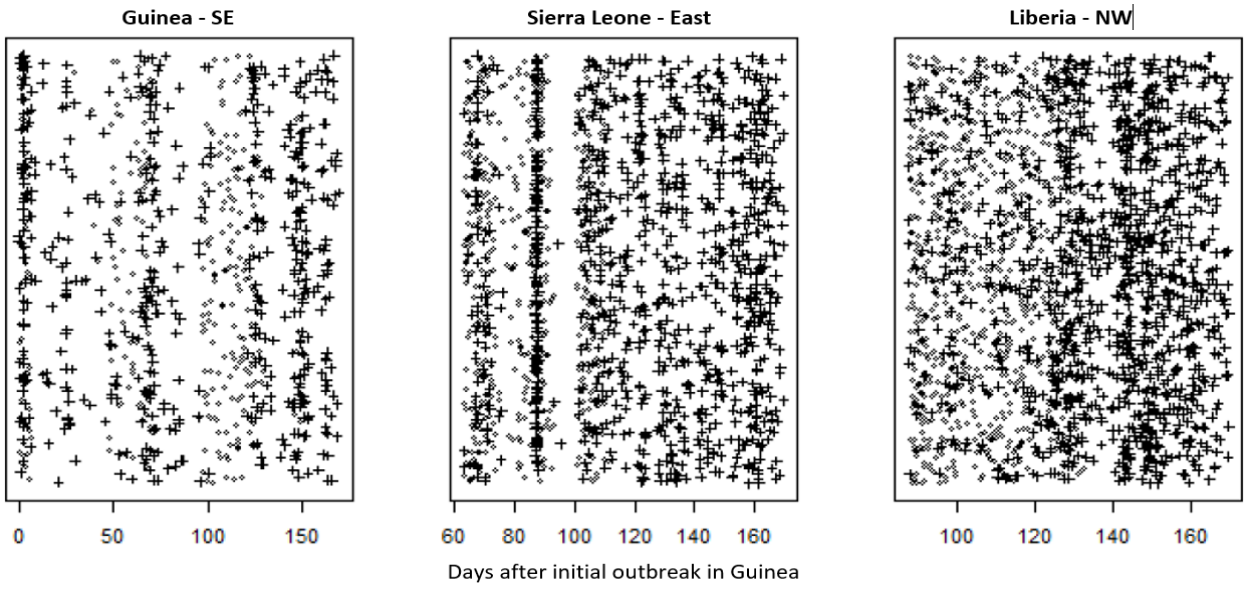
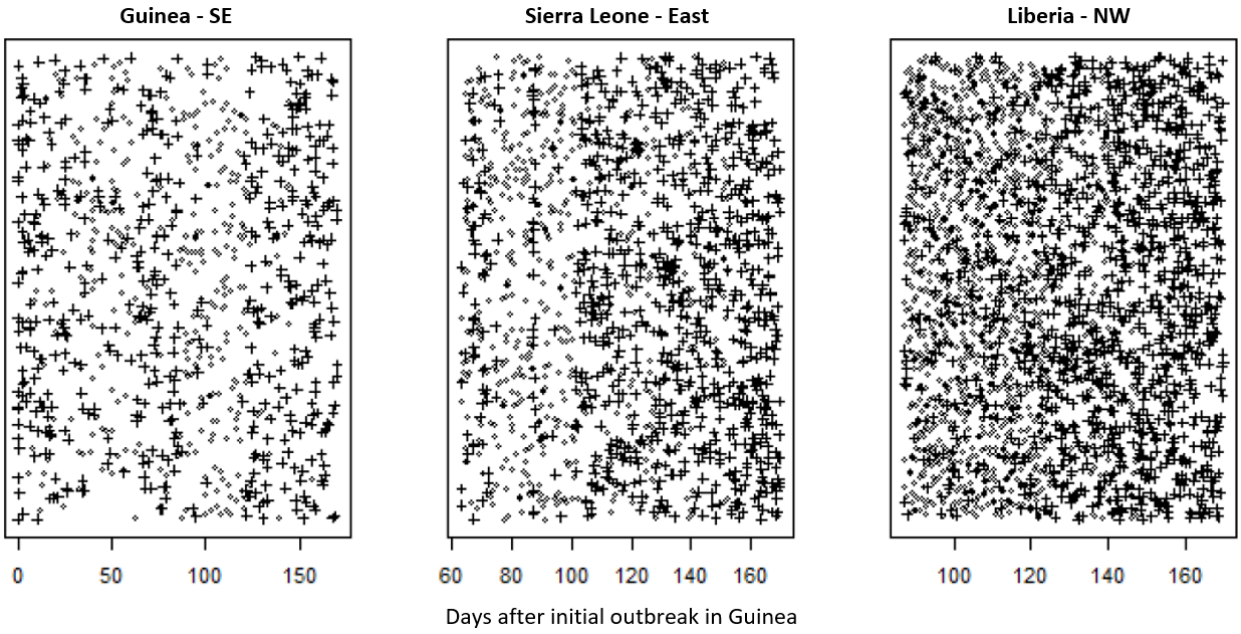


Figure 7. Super-thinning using Hawkes infection rate parameter



Discussion

The results indicate that Hawkes modeling has the potential to perform as good or better than traditional compartmental models such as SEIR in explaining the progression of Ebola disease outbreaks. Hawkes seemed to perform at equal or better levels than SEIR in all aspects of fitting and evaluation. In the prospective analysis, the higher variability of Hawkes simulation estimates may be considered a benefit as it allowed many simulations to line up well with the case load and trajectory of the actual outbreak. However, SEIR modeling also provided some reasonable projection estimates, especially when it was fit to less data.

A weakness to consider in this study is the data itself which is based on official WHO reports. The data likely not comprehensive in accounting for every case of Ebola at the correct time due to limits on human resources in managing the large area and population of the three study regions. Moreover, the data used in evaluation is pseudo data that may not accurately represent the true incidence of cases between reporting dates. Because SEIR modeling is heavily

dependent on the current population of infected individuals, it will be more prone to error if the data is not accurate. With improved detection and reporting measures, as well as more frequent measurements, it is possible that SEIR models could show some improvement and should not be ruled out as useful modeling tools.

Another weakness to consider is that the Hawkes model itself relies more on a background rate and density function than a multiplicative triggering of new infections because each new infection cannot trigger more than 1 new infection. This concept may be somewhat counterintuitive in infectious disease modeling because each infection can generate multiple secondary infections. This may lead to difficultly in properly modeling in the first few weeks of a highly contagious outbreak when the total number of cases is low and should warrant further study. Some modifications to Hawkes have been proposed to account for this issue and shows promising results (Schoenberg et al., 2017). Nonetheless, the results in this paper demonstrate that basic Hawkes modeling can be effective in modeling caseloads several weeks after the outbreak has taken hold.

Conclusion

The analysis indicates that both Hawkes and SEIR models achieve some level of success in describing the spread of the 2014 West Africa Ebola outbreak. Further study should be done to compare these models to other Ebola outbreaks and perhaps data from other infectious diseases. In this data set, Hawkes modeling appeared to provide more accurate results, especially in capturing the variability of long-duration projections. The projections using simulations withholding 25% of data provide the clearest evidence of this accuracy, as the Hawkes simulation variability consistently captured each country's outbreak data. The SEIR model simulation variability, by contrast, was occasionally accurate but in most cases produced

trajectories that were not consistent with the actual data. In future outbreaks, epidemiologists should take great care in using SEIR models to project infections, as the models may poorly predict actual caseloads. They should strongly consider including Hawkes modeling as an alternative or supplement to SEIR models to achieve more well-informed caseload estimates. This in turn could lead to more efficient and effective allocation of resources when managing future outbreaks.

It is important to keep in mind that the spread of Ebola in west Africa in 2014 is one case study that demonstrates the effectiveness of Hawkes modeling. SEIR and Hawkes models may perform differently for other diseases, regions, or time periods. Important subjects for future work would be to compare the fit of Hawkes and SEIR models to data on other diseases and in other regions, and to perform prospective analyses to evaluate the forecasting performance of the two types of models. Such work could help determine if Hawkes modeling is generalizable to explaining future outbreaks.

Another noteworthy consideration for future study is that Hawkes may be expanded to incorporate spatial distribution of cases in the future if the data is available. On the other hand, compartmental modeling is generally limited due to its assumption of spatial homogeneity of each compartment's population. Although some attempts have been made toward spatial compartmental modeling, such as Guofo et al. (2014) who propose a fractional SEIR model using separate S, E, I, and R compartments for each neighboring major metropolitan region in New Zealand with additional terms for the spread between these regions, such models still spatially aggregate the observations resulting in the loss of some information and resolution compared with spatial point process models such as Hawkes models. This advantage to Hawkes models should be explored further to improve its use as a tool in explaining disease outbreak.

Appendix

Data used from Ebola outbreak:

Date	Guinea_SE_Cases	Guinea_SE_Death	SierraLeone_E_Cases	SierraLeone_E_Death	Liberia_NW_Cases	Liberia_NW_Death
23-Mar-14	49	29				
24-Mar-14	86	59				
25-Mar-14	86	60				
26-Mar-14	86	62				
27-Mar-14	103	66				
28-Mar-14	112	70				
31-Mar-14	122	80				
1-Apr-14	127	83				
5-Apr-14	143	86				
7-Apr-14	151	95				
9-Apr-14	158	101				
14-Apr-14	168	108				
16-Apr-14	197	122				
17-Apr-14	203	129				
20-Apr-14	208	136				
23-Apr-14	208	136				
26-Apr-14	224	143				
3-May-14	231	155				
5-May-14	235	157				
6-May-14	236	158				
10-May-14	233	157				
12-May-14	248	171				
23-May-14	258	174				
27-May-14	281	186	16	5		
28-May-14	291	193	50	6		
1-Jun-14	328	208	79	6		
3-Jun-14	344	215	81	6		
5-Jun-14	351	226	89	7		
16-Jun-14	398	264				
17-Jun-14			97	30		
17-Jun-14	390	267	136	55	33	24
19-Jun-14					41	25
20-Jun-14	390	270	136	59		
22-Jun-14					51	34
30-Jun-14	413	303	239	99	107	65
2-Jul-14	412	305	252	101	115	75
6-Jul-14	408	307	305	127	131	84
8-Jul-14	409	309	337	142	142	88
12-Jul-14	406	304	386	194	172	105
14-Jul-14	411	310	397	197	174	106
17-Jul-14	410	310	442	206	196	116
20-Jul-14	415	314	454	219	224	127
23-Jul-14	427	319	525	224	249	129
27-Jul-14	460	339	533	233	329	156
1-Aug-14	485	358	646	273	468	255
4-Aug-14	495	363	691	286	516	282
6-Aug-14	495	367	717	298	554	294
9-Aug-14	506	373	730	315	599	323
11-Aug-14	510	377	783	334	670	355
13-Aug-14	519	380	810	348	786	413
16-Aug-14	543	394	848	365	834	466
18-Aug-14	579	396	907	374	972	576
20-Aug-14	607	406	910	392	1082	624
26-Aug-14	648	430	1026	422	1378	694
31-Aug-14	771	494	1216	476	1698	871
7-Sep-14	861	557	1424	524	2081	1137

References

1. Althaus C.L. (2014). Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLOS Current Outbreaks*.
2. Becker, N. (1977). Estimation for discrete time branching processes with application to epidemics. *Biometrics*, 33(3): 515-522.
3. Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225(1): 24-35.
4. Cao Y., Gillespie D.T., and Petzold L.R. (2007). Adaptive explicit-implicit tau-leaping method with automatic tau selection. *J Chem Phys*, 126(22): 224101.
5. Chowell G., Hengartner N.W., Castillo-Chavez C, Fenimore P.W., and Hyman J.M. (2004). The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J Theor Biol*, 229(1): 119-126.
6. Clements, R.A., Schoenberg, F.P., and Veen, A. (2013). Evaluation of space-time point process models using super-thinning. *Environmetrics*, 23(7), 606-616.
7. Daley, D., and Vere-Jones, D. (2003). An Introduction to the Theory of Point Processes (2^{nd} ed.), New York: Springer.
8. Daley, D., and Vere-Jones, D. (2007). An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure, New York: Springer.
9. Farrington, C.P., Kanaan, M.N., and Gay, N.J. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4(2): 279-295.
10. Guofo, E.F.D., Noutchie, S.C.O., Mugisha, S. (2014). A Fractional SEIR Epidemic Model for Spatial and Temporal Spread of Measles in Metapopulations. *Abstract and Applied*

Analysis, 781028: 1-6.

11. Hawkes, A.G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika,* 58: 83-90.
12. Kermack, W.O. and McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115(772): 700-721.
13. Lekone P.E., and Finkenstädt B.F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170– 7.
14. Liberia Institute of Statistics and Geo-Information Services (LISGIS) (2009). 2008 Population and Housing Census: Final Results. Web. Accessed August 31, 2017.
15. Marsan, D., and Lengliné, O. (2008). Extending earthquakes' reach through cascading. *Science*, 319: 1076-1079.
16. Meyer, L.A. (2007). Contact network epidemiology: bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44(1): 63-86.
17. National Institute of Statistics (2015). General Population and Housing Census (Final Results). Web. Accessed August 31, 2017.
18. Nelder, J.A., and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4): 308-313
19. Ogata, Y. (1978). The asymptotic behavior of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.*, 30(Part A): 243-261
20. Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.,* 83: 9-27.
21. Ogata, Y. (1998). Space-Time Point-Process Models for Earthquake Occurrences. *Annals of the Institute of Statistical Mathematics,* 50: 379-402.

22. Schoenberg, F.P., Hoffman, M., Harrigan, R. (2017). A recursive point process model for infectious diseases. *AISM*, subm. May 22, 2017.
23. Sierra Leone Statistics (2016). Sierra Leone 2015 Population and Housing Census: Provisional Results. Web. Accessed August 31, 2017.
24. Spengler, J.R., Ervin, E.D., Towner, J.S., Rollin, P.E., and Nichol, S.T. (2016). Perspectives on West Africa Ebola Virus Disease Outbreak, 2013-2016. *Emerg Infect Dis.*, 22(6): 956-963.
25. United Nations Development Programme (2015). West African economies feeling ripple effects of Ebola, says UN. Web. Accessed September 14, 2017.
26. WHO Ebola Response Team (2014). Ebola Virus Disease in West Africa – The First 9 Months of the Epidemic and Forward Projections. *N Engl J Med*, 371: 1481-1495.
27. World Health Organization (2003). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. Web. Accessed September 2, 2017.
28. World Health Organization (2016). Ebola data and statistics. Web. Accessed August 31, 2017.