# A non-parametric Hawkes model of the spread of Ebola in West Africa

Junhyung Park, Adam W. Chaffee, Ryan J. Harrigan,
Frederic Paik Schoenberg

*University of California, Los Angeles*

December 2, 2018

**Abstract**

Recently developed methods for the non-parametric estimation of Hawkes point process models facilitate their application for describing and forecasting of spread of epidemic diseases. We use data from the 2014 Ebola outbreak in West Africa to evaluate how well a simple Hawkes point process model can forecast the spread of Ebola virus in Guinea, Sierra Leone, and Liberia. For comparison, SEIR models fit previously to the same data are evaluated using identical metrics. To test the predictive power of each of the models, we simulate the ability to make near real-time predictions during an actual outbreak by using the first 75% of the data for estimation and the subsequent 25% of the data for evaluation. Hawkes models accurately describe the spread of Ebola in each of the three countries investigated, and result in a 38% reduction in RMSE for weekly case estimation across all countries when compared to SEIR models (total RMSE of 59.8 cases/week using SEIR compared to 37.1 for Hawkes). We demonstrate that the improved fit from Hawkes modeling cannot be attributed to overfitting alone, and evaluate the advantages and disadvantages of Hawkes models in general for forecasting the spread of epidemic diseases.

*Keywords:* Compartmental models, disease epidemics, non-parametric estimation, point processes, SEIR models, self-exciting.

# 1 Introduction

Between March 2014 and June 2016, the West African countries of Guinea, Sierra Leone, and Liberia experienced a historical Ebola outbreak, one that eventually surpassed all other previous Ebola outbreaks combined in terms of total cases reported (WHO Ebola Response Team, 2014). The epidemic resulted in nearly 30,000 infections and more than 11,000 deaths (WHO 2016), and also took a severe toll on the economy and quality of life in the region, due to decreased trade, border closures, and decreased foreign investment (United Nations Development Programme, 2015). To mitigate future outbreaks of highly infectious diseases, it is important that governments focus on improving detection and response capacity, among other important public health policy objectives (Spengler et al., 2016). To this aim, statistical models can play an important role in forecasting the spread of infectious diseases both during and after an outbreak, leading to more effective allocation and mobilization of public health 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 SIR (Susceptible-Infected-Recovered) model and its variants. Such models involve classifying individuals according to disease status, and then modeling the changes in numbers of infected, susceptible, and recovered individuals in each group using a series of differential equation models. Compartmental modeling has grown to become a primary resource of the epidemiological community for modeling the establishment and spread of many infections such as HIV, SARS, and influenza. In recent decades, traditional SIR models have been modified with new parameters or more informed parameter estimates, to better fit individual disease characteristics (Britton, 2010).

Among these derivations, 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 the recent West African Ebola outbreak by Althaus (2014). While effective in predicting some aspects of outbreaks, SEIR models rely heavily on accurate parameter estimation and rely critically on the assumption of no community intervention in response to outbreaks as well as the mass action assumption that all susceptible members of the population are equally likely to be infected (Chowell et al. 2004, Meyers 2007). Although departures from these assumptions are common and frequently result in wildly inaccurate forecasts, these models provided important insights about the potential progression of the disease without intervention, and played a critical role in eliciting a swift public health response.

We put forth a simple, non-parametrically estimated Hawkes point process model as an addition to the many popular methods used in the disease modeling and forecasting toolbox. Hawkes models are currently widely used in seismology to describe earthquake catalogs. Though these models have outperformed their competitors in earthquake forecasting experiments (Schorlemmer et al. 2010, Zechar et al. 2013), rarely have they been applied to the emergence and spread of infectious diseases. Self-exciting point 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. Recently, Balderama et al. (2012) fit a modified Hawkes model to sightings of one invasive species of red banana trees spreading in a Costa Rican rainforest, and Meyer et al. (2012) used a parametric Hawkes model for the incidence of invasive meningococcal disease (IMD) in humans, and the results proved useful for estimating spread rates and for the detailed description of properties of the outbreak. However, both Balderama et al. (2012) and Meyer et al. (2012) relied on parametric forms for the triggering function rather than more general non-parametric methods, and neither compared the fit or performance of their fitted model with those of more traditional compartmental models.

Here, for the first time, we compare the performance of non-parametrically estimated Hawkes point process models to more traditional compartmental models (SEIR) for estimating the spread of an infectious disease outbreak. While many variants and advances exist for SEIR modeling, including time-varying transmission rates and additional compartments for pending funerals, hospitals and exposed health care workers (see Viboud et al. 2018, Funk et al. 2018, Champredon et al. 2018), we sought to test the most general, least parameterized versions of each model class to enable a baseline-to-baseline comparison. Fortunately, the application to the spread of Ebola in West Africa by Althaus (2014) provides an ideal test case, where simple SEIR models have already been proposed, fitted, and shown by an expert to provide a good description of an outbreak in a well-vetted, large dataset with replicates across varying environmental and economic conditions in Guinea, Liberia and Sierra Leone.

Comparisons between point process models and compartmental models are particularly illuminating, as the two types of models rely on different assumptions and operate under fundamentally different mathematical frameworks. As such, Hawkes models may provide different insights into the spread of epidemics and invasive species as compared to more typical models, including a description of the spread via an estimated triggering kernel.

The structure of this paper is as follows. Following a description of the data in Section 2, we briefly review Hawkes and SEIR models in Section 3, as well as methods for model fitting and assessing their fit. In Section 4, we compare the fit of the two models in Guinea, Sierra Leone, and Liberia. A discussion and some concluding remarks are given in Section 5.

# 2 Data

Data were collected and aggregated from the World Health Organization (WHO) outbreak reports on Ebola during and after the outbreak period (WHO, 2016). These reports are typically released sub-weekly by WHO and include the country, geographic location within country (either by region, closest city, or village) as well as confirmed cases and deaths from Ebola virus. Following Althaus (2014), data were filtered to include only a count of infection cases from Ebola at regular, reported time points in three regions: Southeast Guinea, Eastern Sierra Leone, and Northwest Liberia. The time range of these observations begins on 2014-03-23 and ends on 2014-09-07, again to align with Althaus (2014). A copy of the data used can be found at http://wildfire.stat.ucla.edu/rick/ebola. In fitting Hawkes models, estimated occurrence times were distributed uniformly within report dates. For a small number of report dates, the cumulative count of cases was subsequently revised downwards by WHO; these revisions are ignored in the current analysis, as in Althaus (2014). The cumulative count of cases reported by the WHO, and the data used to fit Hawkes models are plotted in Figure 1.

# 3 Methods

## 3.1 Hawkes models and their non-parametric estimation.

A temporal point process is a collection of points occurring on the real line. Its associated counting process $N(t)$ counts the number of points occurring between time 0 and time $t$, inclusive. Such a process is usually characterized via its *conditional intensity* $\lambda(t)$, which is the infinitesimal expected rate at which points are accumulating at time $t$, given the history $\mathcal{H}_t$ of all points occurring prior to time $t$ (Daley & Vere-Jones, 2003):

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

Because the conditional intensity is a function of time, it is convenient for describing events that may happen with rates that change dynamically. Hawkes models are often extended to data occurring in both space and time. However, to provide an equivalent comparison to SEIR modeling where cases are aggregated over an entire spatial region as in Althaus (2014), we consider a purely temporal Hawkes process (Hawkes 1971), where $\lambda(t)$ is written as:

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

This is sometimes called a *branching* process, or *epidemic* by Ogata (1988) because every occurrence $t_i$ contributes a secondary series of occurrences ($a$ftershocks) occurring at a time-varying rate $Kg(t - t_i)$, which in turn produces its own aftershock sequence, and so on.

Typically the triggering function $g$ in equation (1) is constrained to be a density, so the parameter $K$ represents the expected number of new infections directly attributable to each case. Since each case thus is posited to cause an expected number $K$ of secondary infections, any particular case is expected to be an ancestor to $K + K^2 + K^3 + ... = \frac{1}{1-K} - 1$ total infections. Thus $K$ should satisfy $0 \leq K < 1$ in order for the process to be stable.

For many processes, the triggering density $g(u)$ decays gradually as the time delay $u$ increases. Model fitting typically involves choosing a parametric form for $g(u)$, and maximizing the log-likelihood function (Daley & Vere-Jones, 2003),

$$l(\theta) = \sum_k \log[\lambda(t_k; \theta)] - \int_0^T \lambda(t; \theta) du \tag{2}$$

where $\theta$ is the vector of parameters governing the shape of $g$ and $[0, T]$ is the time window of observation.

In applications to seismology, the parametric form of $g$ has been well established over decades of research. Because infectious diseases are a relatively newer application, it is desir-

able to make no assumption about the shape of $g$. Therefore, we choose to non-parametrically estimate the triggering function $g$, constant background rate $\mu$ and productivity constant $K$ using the method proposed by Marsan and Lengliné (2008). This method assumes that $g$ is a piecewise constant step function with user-defined number of steps and unknown heights estimated by approximate maximum likelihood. A key principle driving this methodology is that, given a model for $\lambda(t)$, the probability that a point $t_j$ occurred exogenously due to the background rate is

$$p_{jj} = \frac{\mu}{\lambda(t_j)}, \tag{3}$$

and the probability that point $t_i$ triggered point $t_j$ is

$$p_{ij} = \frac{Kg(t_j - t_i)}{\lambda(t_j)}, \tag{4}$$

as noted in Zhuang, Ogata and Vere-Jones (2002). An initial guess, $p_{jj}^{(0)}$ and $p_{ij}^{(0)}$ for $1 \leq i < j \leq N(T)$, gives the full probabilistic branching structure of the point process, and this is used to obtain estimates $\mu^{(0)}$, $K^{(0)}$, and the steps heights of $g(u)^{(0)}$ following the non-parametric procedure described in Marsan and Lengliné (2008) and Fox et al. (2016). Knowing these in turn allows updated probabilities $p_{jj}^{(1)}$ and $p_{ij}^{(1)}$ to be computed using (3) and (4). This is iterated until the largest update in any $p_{jj}^{(k)}$ or $p_{ij}^{(k)}$ is less than some small constant $\epsilon$. All parameters and triggering densities are estimated separately for each region.

## 3.2 SEIR models and their estimation.

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):

$$\frac{dS}{dt} = -\beta(t)\frac{SI}{N} \tag{5}$$

$$\frac{dE}{dt} = \beta(t)\frac{SI}{N} - \sigma E \tag{6}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{7}$$

$$\frac{dR}{dt} = \gamma I. \tag{8}$$

Here $S$ is the size of the susceptible population, $E$ is the size of the population that has contracted Ebola but is not yet infectious (latent population), $I$ is the size of the infectious population, and $R$ is the size of the recovered/deceased population. These four quantities sum to $N$, the total population. The populations are assumed to be spatially aggregated, 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 $\kappa$:

$$\beta(t) = \beta e^{-\kappa t}, \tag{9}$$

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, $\sigma$, and the rate of death or recovery, $\gamma$. In model fitting, $\gamma$ and $\sigma$ are typically assumed constant, as in Althaus (2014).

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

large proportion of the population. When $R_0(t)$ drops below 1, the epidemic is unsustainable (Diekmann and Heesterbek 2000, Lipsitch et al. 2003).

As in Althaus (2014), parameter estimates for SEIR models were obtained using maximum likelihood estimation (MLE) assuming that occurrences of new cases follow a Poisson distribution. SEIR models were fit separately for Guinea, Sierra Leone, and Liberia using the original discretely reported Ebola outbreak data containing only cases and deaths at each reporting date, as in Althaus (2014). The theoretical SEIR model outlined above is purely deterministic, so to convert this process into a stochastic model for simulating real-world outbreaks, and to facilitate model evaluation using statistical methods, the tau-leaping approximation of Cao et al. (2007) was applied. Under the Tau-leaping simulation method applied to SEIR, transitions from each of the S, E, I, and R populations are simulated based on a Markov-Chain Monte Carlo (MCMC) process. Transition probabilities from S to E, E to I, and I to R are all calculated using the current state populations and the fitted $R_0(t)$, $\kappa$, $\sigma$, and $\gamma$. Under each round of simulation, at random one person at one time point will transfer to a new population based on these probabilities. The transition probabilities are then recalculated based on the new state populations. The process continues until the end of the observed time window is reached. Optimization and estimation of the SEIR model were performed using a Nelder-Mead algorithm and the deSolve *R* package (Soetaert et al., 2010), and the tau-leaping method was performed using the adaptivetau *R* package (Johnson, 2016).

## 3.3 Evaluation Techniques

Simulations of both SEIR and Hawkes models were used in order to assess statistically the compatibility of the observations with forecasts made with each model. SEIR and Hawkes parameter estimates obtained using all available data for each country were used for each

of 1,000 simulations, and the mean of simulations for each week and each country were recorded and compared to the actual number of infections per week. Here, different scoring rules are possible as surveyed in Gneiting and Katzfuss (2014), and we focus primarily on the root mean square error (RMSE) of weekly predictions and note that our main findings are not substantially influenced by the choice of scoring rule. SEIR simulations require assumed values for the starting populations on day 0. Hawkes models do not require such assumptions. However, according to the fitted Hawkes model, simulated infections near the beginning of the recording period are assumed to result exclusively from the background process rather than contagion, which may lead to under-predictions in week 1. Since both models are impractical for estimating cases in week 1, only estimates from week 2 and beyond are used for comparison. The susceptible population was set based on the most recently published census data from Guinea (National Institute of Statistics, 2015), Sierra Leone (Sierra Leone Statistics, 2016), and Liberia (LISGIS, 2009), under the assumption that approximately the entire living population is susceptible. The infectious populations were set based on the observed data from the WHO.

In a separate analysis to account for the possibility of over-fitting in a retrospective analysis such as this, and to assess the ability of the models to forecast spread during an actual outbreak, parameter estimates were fitted using only the infections occurring in the first 75% of the observation window, and the resulting fitted models were then used to project cumulative infections for the remaining period. This analysis also involved 1,000 simulations per country for both SEIR and Hawkes.

Because the Hawkes and the SEIR model with tau-leaping rely on different underlying probabilistic processes, the likelihood of a given model is difficult to interpret substantively, but the relative fit of two distinct models can readily be compared using log-likelihoods. For nested point process models, the difference in log-likelihood 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). For non-nested models, a common alternative in maximum likelihood estimation is to calculate the Akaike Information Criterion (AIC), which is calculated as $-2\cdot$log-likelihood $+ 2p$ where $p$ is the number of estimated parameters in the model (Ogata, 1988). Lower AIC indicates better fit.

The competing SEIR and Hawkes models were also evaluated using superthinning (Clements et al. 2013). In superthinning, the existing data points are first thinned where each point is randomly kept independently of the others with probability $\min\{b/(\hat{\lambda}(t)), 1\}$, and then new points are superposed according to a Poisson process over the observed time window $[0, T]$, with rate $(b-\hat{\lambda}(t))^+$. Superthinning requires an initial choice of the tuning parameter, $b$, and as suggested in Clements et al. (2013), we used the simple default value of the total number of cases divided by the length, in days, of the observation period. For the SEIR model, the value of $\hat{\beta}(t)$ multiplied by the size of the infectious population at time $t$ was used as the estimated rate function $\hat{\lambda}(t)$ to calculate thinning and superposing probabilities. After superthinning, the resulting residual process is a homogeneous Poisson process with rate $b$ if and only if the estimate of the conditional intensity, $\hat{\lambda}$, is correct (Clements et al. 2013), and thus one may examine the superthinned residuals for uniformity. Sparsity of points in the superthinned residuals corresponds to areas where the model over-predicted, whereas clustering in the superthinned residuals indicates areas where the model under-predicted the number of observed cases.

# 4 Results

## 4.1 Model Fitting and Weekly Estimates

Figure 2 displays non-parametric estimates of the Hawkes triggering density in (1) for each country. According to the fitted model, in Guinea and Liberia, an infected individual directly triggered new infections on the scale of up to 3 days. In Sierra Leone, this density appeared to decay somewhat faster, with most triggering occurring within 1 day, according to the fitted Hawkes model. Note that the estimated triggering times in Figure 2 are times between *recorded diagnoses* of cases of Ebola; such recordings may be substantially more clustered than actual transmissions of the disease.

Table 1 shows the estimated parameters of the Hawkes model (1) for each country. More intense clustering was observed in Liberia and a significantly smaller percentage of the points are attributed to the background rate according to the fitted Hawkes model. According to the fitted Hawkes model (1), 89% of cases in Guinea were attributable to contagion from other observed cases in Guinea, whereas in Sierra Leone an estimated 93% of cases were attributed to contagion from other observed cases, and in Liberia the corresponding estimate was 99%.

The log-likelihood scores for all models are shown in Table 2. The AIC for the Hawkes model is lower than that of the SEIR model for all three countries, indicating that the Hawkes models provided a better fit to the infection outbreak data in all three countries. Hawkes models also had correspondingly lower RMSE in weekly predictions compared to SEIR models, for all three countries. The total RMSE across all countries combined was 59.8 cases/week using SEIR and 37.1 cases/week using the Hawkes model (1), which represents a 38% decrease in the RMSE.

Weekly estimates of total infections per week based on the mean of 1000 simulations of the Hawkes and SEIR models are displayed along with the observed number of infections in

Figure 3. The weekly simulation means for the fitted SEIR model resemble a lagged version of the observed weekly counts with a lag of two weeks.

Hawkes models show a similar dependence on past fluctuations in cases due to the cascading nature of Hawkes processes; a large number of cases in the previous week will be expected to trigger more simulated infections into the following week. However, the week to week dependence is weaker and more complex than in the fitted SEIR model. For example, in weeks 18 through 20 in Guinea SE, the Hawkes weekly projection behaves like a one week lag of the actual data. At other times, however, the Hawkes estimate appears to be more adaptive than a mere one week lag, such as in forecasting the sudden decrease in infections from week 11 to 12 in Guinea SE, or the rise in infections from week 5 to 7 in Sierra Leone East. Across all three countries the Hawkes models tend to produce more accurate weekly estimates with less extreme errors.

The errors in weekly projections for the Hawkes and SEIR models are displayed in Figure 4. The errors for SEIR tend to be slightly more variable than those for the Hawkes model. One sees in Figure 4 that the Hawkes model tends to have smaller prediction errors than the SEIR model for weekly projections of numbers of new infections, and this improved performance seems to persist in all three West African countries and throughout the duration of the outbreak.

## 4.2 Prospective Out-of-Sample Prediction

Figure 5 shows simulations of SEIR and Hawkes models fitted using only the first 75% of data for each outbreak and simulated for the remaining 25% of the time. For Sierra Leone, the SEIR model simulations forecast the trajectory of the number of new infections quite accurately, especially during the first 16 days of the simulations. The SEIR model significantly underestimated the number of new infections in Guinea throughout the course

of the simulation. For Liberia, the simulations of the fitted SEIR model initially tended to overestimate the number of observed infections, due to the approximately exponential predicted acceleration characteristic of the SEIR model.

Simulated Hawkes processes for Guinea were remarkably accurate for the first 20 days of the simulations, and tended on average to underpredict the number of new infections after 30 days. The simulations of the Hawkes model in Sierra Leone and Liberia also tended to slightly underpredict the number of infections. In all three countries, the variation in the Hawkes simulations was much greater than that of the SEIR model. Using the average of the simulations as a forecast, the total RMSE of the forecasts in the first two weeks for SEIR was 208.5 cases/day, compared to 60.5 cases/day for the Hawkes model, representing a 71% reduction in error. The results suggest that overfitting is not responsible for the improved performance of the Hawkes model for describing the spread of Ebola in these 3 countries, since in these comparisons the models were assessed using data not used in the parameter estimation.

## 4.3 Superthinning Analysis

Superthinning results are displayed in Figure 6 for all three regions. The superthinned residuals corresponding to the SEIR model for Guinea have obvious clustering during the first week and around 2014-06-01 and 2014-08-20, indicating times when the model underestimated the number of new infections. This likely occurs because the estimated conditional intensity is heavily dependent on the current infectious population which can be highly variable. Immediately after an unexpected surge in observed infections, the modeled rate according to the SEIR model tends to remain relatively low for approximately 2 weeks, and as a result, most of the observed points are retained after superthinning, resulting in intense clustering in the residuals. The SEIR superthinned residuals for Guinea also have noticeable gaps

and regions of sparsity of points, particularly around 2014-07-01, corresponding to the SEIR model overestimating the number of new cases. The superthinned residuals corresponding to Sierra Leone indicate poor fit in many places as well and most notably underestimation of the number of new infections around 2014-05-21, as well as overestimation around 2014-07-01. The superthinned residuals for Liberia corresponding to the SEIR model show few noticeable departures from uniformity despite some sparsity of points around 2014-08-05.

The superthinned residuals corresponding to the fitted Hawkes models, shown in Figure 6, show no substantial departures from homogeneity for all three countries. The Hawkes model does not have any gross inaccuracies in Guinea, Sierra Leone, or Liberia and appears to more closely describe the spread of new infections in these three countries.

# 5 Discussion

The application of non-parametric Hawkes point process models to predict the spread of Ebola virus in West Africa indicates that these novel methods have the potential to be a useful addition to the pallet of available methods for disease forecasting. In all aspects of fitting and evaluation of Ebola spread that we performed, Hawkes models performed as well as, or better than, SEIR models. Our results do not suggest traditional SEIR models should be discounted; rather, they highlight the utility of Hawkes processes as an alternative and novel framework with which to predict disease spread to reveal new insights to how outbreaks may evolve. SEIR and Hawkes models provide two very different descriptions of an outbreak and focus on different aspects of the disease spread. Therefore, we suggest that ensemble forecasts may provide maximum insights to inform public health decision making.

For instance, in weekly forecasts, Hawkes models appeared to provide more accurate predictions than SEIR. This is possibly because a Hawkes model, unlike SEIR, works by

estimating the temporal distribution of times between *individual* reported infections via the triggering kernel. This provides new information about the dynamics of an outbreak that is highly localized in time.

One might object that, in a retrospective analysis such as this, the improvement in fit might be due to fitting a more complex model with more free parameters, in which case overfitting might be a problem and the improvement would be unlikely to be replicated in further applications, particularly if the model were used in forecasting. However, we found that even when models were fit using one portion of the data and assessed on a different portion of the data, simulating performance *during* an outbreak, Hawkes models were more accurate than SEIR by a 71% reduction in RMSE for the first two weeks of forecasting. This suggests that Hawkes models may in the future be used for more accurately forecasting the spread of epidemic diseases, which could help facilitate and inform surveillance and mitigation efforts to help curb outbreak spread as it occurs.

One limitation of our Hawkes models is their reliance on a constant supply of exogenous infections from the background rate for the continued propagation of the disease. According to the fitted model, subsequent triggered infections eventually die out since each infection only directly triggers a Poisson($K$) number of new infections where $K < 1$. Such a Hawkes model would fit poorly to data exhibiting intense clustering of observed cases. This may also lead to difficulty in properly modeling the first few weeks of a contagious outbreak when the total cumulative number of cases is low or the latter parts of an outbreak where disease spread is nearly saturated and slowing down due to human containment or intervention. Some modifications to Hawkes models have been proposed to account for this issue and show promising results (Schoenberg et al., 2019). Nonetheless, the results in this paper demonstrate that basic Hawkes modeling can be effective in modeling caseloads up to a few weeks after an outbreak has emerged.

It should be noted, however, that the WHO data considered here consist of periodically updating counts of cases (on average 4 days between updates). A higher temporal resolution will be particularly desirable for future research in order to improve the accuracy and assessment of point process models. The data are 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. Because SEIR modeling is heavily dependent on the current population of infected and susceptible individuals, it may be hypothesized to be more sensitive to missing data and errors in reporting, though this should be studied and quantified in future research. While 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, and time periods. It is recommended that future investigations 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 whether Hawkes modeling will provide a generalizable framework for prediction across a variety of infectious diseases, with highly disparate outbreak periods.

Another important area for future study is to use spatial-temporal triggering densities in Hawkes models, to describe the detailed spatial-temporal distribution of infections when sufficient spatial precision is available. Whether classic exponential and power law kernels work just as well as the non-parametric approach can be checked as well. In this paper, we limited the Hawkes model to purely temporal triggering in each spatial region, in order for the Hawkes and SEIR models to be comparable and so both models could be estimated using the data of Althaus (2014). However, one advantage of Hawkes models is their natural generalization to the case of further spatial precision. By contrast, compartmental modeling is generally limited due to its assumption of spatial homogeneity of each compartment's pop-

ulation. Some attempts have been made in this regard through meta-population modeling and spatial compartmental modeling by Keeling & Rohani (2007) and Guofo et al. (2014), respectively. The latter, for example, 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-temporal point process models such as Hawkes models. These considerations should provide impetus for future model refinement and improvements for each model class that are likely to further improve our understanding of the evolutionary nuances of disease outbreaks.

## Acknowledgements.

This material is based upon work supported by the National Science Foundation under grant number DMS 1513657.

## References

[1] C.L. Althaus. "Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa." In: *PLOS Current Outbreaks* (2014).

[2] E. Balderama et al. "Application of branching point process models to the study of invasive red banana plants in Costa Rica." In: *JASA* 107.498 (2012), pp. 467–476.

[3] N. Becker. "Estimation for discrete time branching processes with application to epidemics." In: *Biometrics* 33.3 (1977), pp. 515–522.

[4] T. Britton. "Stochastic epidemic models: A survey." In: *Mathematical Biosciences* 225.1 (2010), pp. 24–25.

[5] Y. Cao, D.T. Gillespie, and L.R. Petzold. "Adaptive explicit-implicit tau-leaping method with automatic tau selection." In: *J Chem Phys* 126.22 (2010), p. 224101.

[6] D. Champredon et al. "Two approaches to forecast Ebola synthetic epidemics." In: *Epidemics* 22 (2018), pp. 36–42.

[7] G. Chowell et al. "The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda". In: *J Theor Biol* 229.1 (2004), pp. 119–126.

[8] R.A. Clements, F.P. Schoenberg, and A. Veen. "Evaluation of space-time point process models using super-thinning." In: *Environmetrics* 23.7 (2013), pp. 606–616.

[9] D. Daley and D. Vere-Jones. "An Introduction to the Theory of Point Processes. (2nd ed.)" In: New York: Springer (2003).

[10] D. Daley and D. Vere-Jones. "An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure". In: New York: Springer (2007).

[11] O. Diekmann and J.A.P. Heesterbek. "Mathematical epidemiology of infectious diseases: model building, analysis and interpretation". In: *Chichester, UK: Wiley* (2000).

[12] C.P. Farrington, M.N. Kanaan, and N.J. Gay. "Branching process models for surveillance of infectious diseases controlled by mass vaccination." In: *Biostatistics* 4.2 (2003), pp. 279–295.

[13] E.W. Fox, F.P. Schoenberg, and J.S. Gordon. "Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences." In: *Annals of Applied Statistics* 10.3 (2016), pp. 1725–1756.

[14] S. Funk et al. "Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model". In: *Epidemics* 22 (2018), pp. 56–91.

[15] T. Gneiting and M. Katzfuss. "Probabilistic Forecasting". In: *Annual Review of Statistics and Its Applications* 1.1 (2014), pp. 125–151.

[16] E.F.D. Guofo, S.C.O. Noutchie, and S. Mugisha. "A Fractional SEIR Epidemic Model for Spatial and Temporal Spread of Measles in Metapopulations." In: *Abstract and Applied Analysis* 781028 (2014), pp. 1–6.

[17] A.G. Hawkes. "Spectra of some self-exciting and mutually exciting point processes." In: *Biometrika* 58 (1971), pp. 83–90.

[18] L. Hunt et al. "Clinical presentation, biochemical, and haematological parameters and their association with outcome in patients with Ebola virus disease: an observational cohort study." In: *The Lancet Infectious Diseases* 15.11 (2015), pp. 1292 –1299.

[19] Philip Johnson. *adaptivetau: Tau-Leaping Stochastic Simulation*. R package version 2.2-1. 2016. URL: https://CRAN.R-project.org/package=adaptivetau.

[20] Matt J. Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. 1st ed. Princeton, N.J.: Princeton University Press, Oct. 2007. ISBN: 9780691116174. URL: http://amazon.com/o/ASIN/0691116172/.

[21] W.O. Kermack and A.G. McKendrick. "A contribution to the mathematical theory of epidemics." In: *Proceedings of the Royal Society A* 115.771 (1927), pp. 700–721.

[22] P.E. Lekone and B.F. Finkenstädt. "Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study." In: *Biometrics* 62.4 (2006), pp. 1170–7.

[23] M. Lipsitch et al. "Transmission dynamics and control of severe acute respiratory syndrome." In: *Science* 300.5627 (2003), pp. 1966–70.

[24] D. Marsan and O. Lengliné. "Extending earthquakes reach through cascading." In: *Science* 319 (2008), pp. 1076–1079.

[25] S. Meyer, J. Elias, and Höhle M. "A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence." In: *Biometrics* 68 (2012), 607616.

[26] L.A. Meyers. "Contact network epidemiology: bond percolation applied to infectious disease prediction and control." In: *Bulletin of the American Mathematical Society* 44.1 (2007), pp. 63–86.

[27] National Institute of Statistics (Guinea). "General Population and Housing Census (Final Results)". In: (2015), Accessed August 31, 2017. URL: http://www.stat-guinee.org.

[28] J.A. Nelder and R. Mead. "A Simplex Method for Function Minimization." In: *The Computer Journal* 7.4 (1965), pp. 308–313.

[29] Y. Ogata. "Space-Time Point-Process Models for Earthquake Occurrences." In: *Annals of the Institute of Statistical Mathematics* 50 (1998), pp. 379–402.

[30] Y. Ogata. "Statistical models for earthquake occurrences and residual analysis for point processes." In: *J. Amer. Statist. Assoc.* 83 (1988), pp. 9–27.

[31] Y. Ogata. "The asymptotic behavior of maximum likelihood estimators for stationary point processes." In: *Ann. Inst. Statist. Math* 30.Part A (1978), pp. 243–261.

[32] F.P. Schoenberg, M. Hoffman, and R. Harrigan. "A recursive point process model for infectious diseases." In: *Annals of the Institute of Statistical Mathematics* (To Appear) (2019).

[33] D. Schorlemmer et al. "First results of the Regional Earthquake Likelihood Models Experiment." In: *Pure and Applied Geophysics* 167 (2010), pp. 859–876.

[34] Sierra Leone Statistics. "Sierra Leone 2015 Population and Housing Census: Provisional Results." In: (2016), Accessed August 31, 2017. URL: https://www.statistics.sl/wp-content/uploads/2017/01/final-results_-2015_population_and_housing_census.pdf.

[35] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. "Solving Differential Equations in R: Package deSolve". In: *Journal of Statistical Software* 33.9 (2010), pp. 1–25. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i09. URL: http://www.jstatsoft.org/v33/i09.

[36] J.R. Spengler et al. "Perspectives on West Africa Ebola Virus Disease Outbreak." In: *Emerg Infect Dis.* 22.6 (2016), pp. 956–963.

[37] Liberia Institute of Statistics and Geo-Information Services (LISGIS). "Population and Housing Census: Final Results. (2008)". In: (Accessed August 31, 2017). URL: https://www.lisgis.net/others.php?&7d5f44532cbfc489b8db9e12e44eb820=NTM1.

[38] United Nations Development Programme. "West African economies feeling ripple effects of Ebola, says UN." In: (2015), Accessed September 14, 2017. URL: http://www.undp.org/content/undp/en/home/presscenter/pressreleases/2015/03/12/west-african-economies-feeling-ripple-effects-of-ebola-says-un.html.

[39] C. Viboud et al. "The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt". In: *Epidemics* 22 (2018), pp. 13–21.

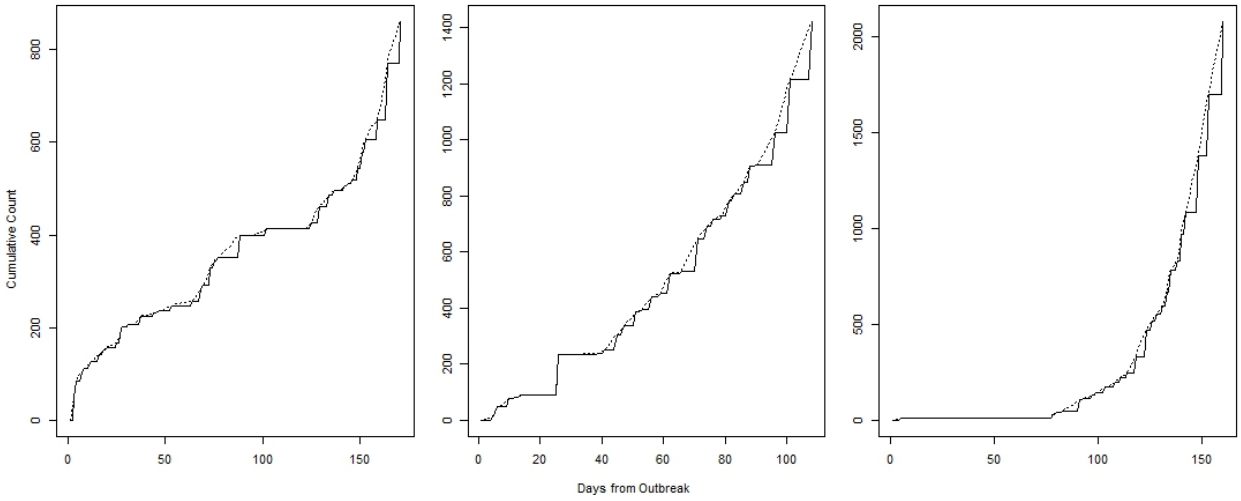
[40] WHO Ebola Response Team. "Ebola Virus Disease in West Africa The First 9 Months of the Epidemic and Forward Projections." In: *N Engl J Med* 371 (2014), pp. 1481–1495.

[41] World Health Organization. "Ebola data and statistics." In: (2016), Accessed August 31, 2017. URL: http://apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-latest?lang=en.

[42] World Health Organization. "Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003." In: (2003), Accessed September 2, 2017. URL: http://www.who.int/csr/sars/country/table2004_04_21/en/.

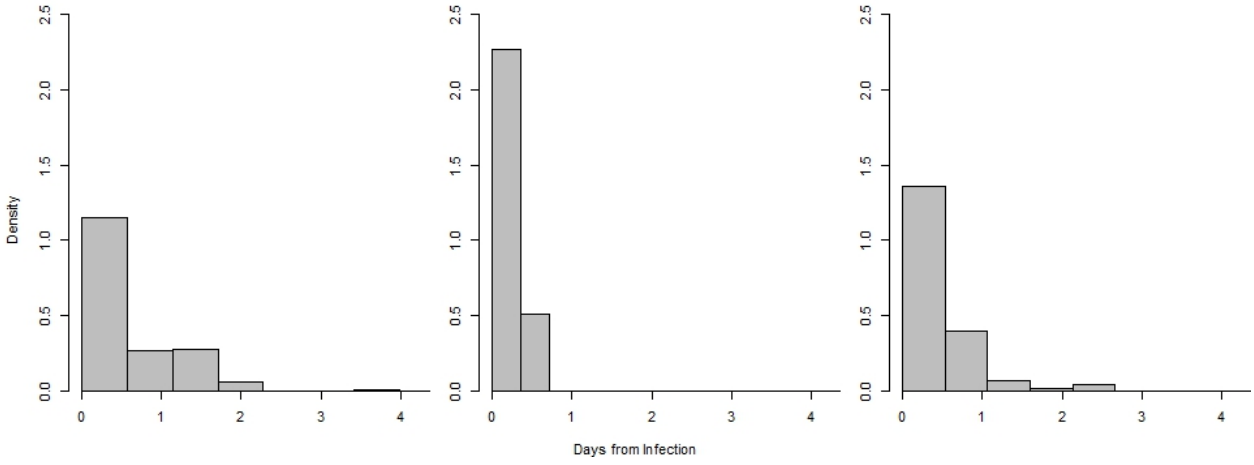
[43] J.D. Zechar et al. "Regional earthquake likelihood models I: First-order results." In: *Bull. Seismol. Soc. Amer.* 103.2A (2013), pp. 787–798.

Figure 1: Point process vs. WHO cumulative case counts




(Left to right): Guinea SE, Sierra Leone East, Liberia NW. Solid = cumulative number of cases reported by WHO, dashed = cumulative number of cases reported by WHO with times uniformly spread within WHO report dates. The start dates of outbreak from left to right are 2014-03-23, 2014-05-27 and 2014-04-05 respectively.

Figure 2: Estimated Hawkes triggering density




(Left to Right): Guinea SE, Sierra Leone East, Liberia NW.

Table 1: Hawkes parameter estimates and standard errors

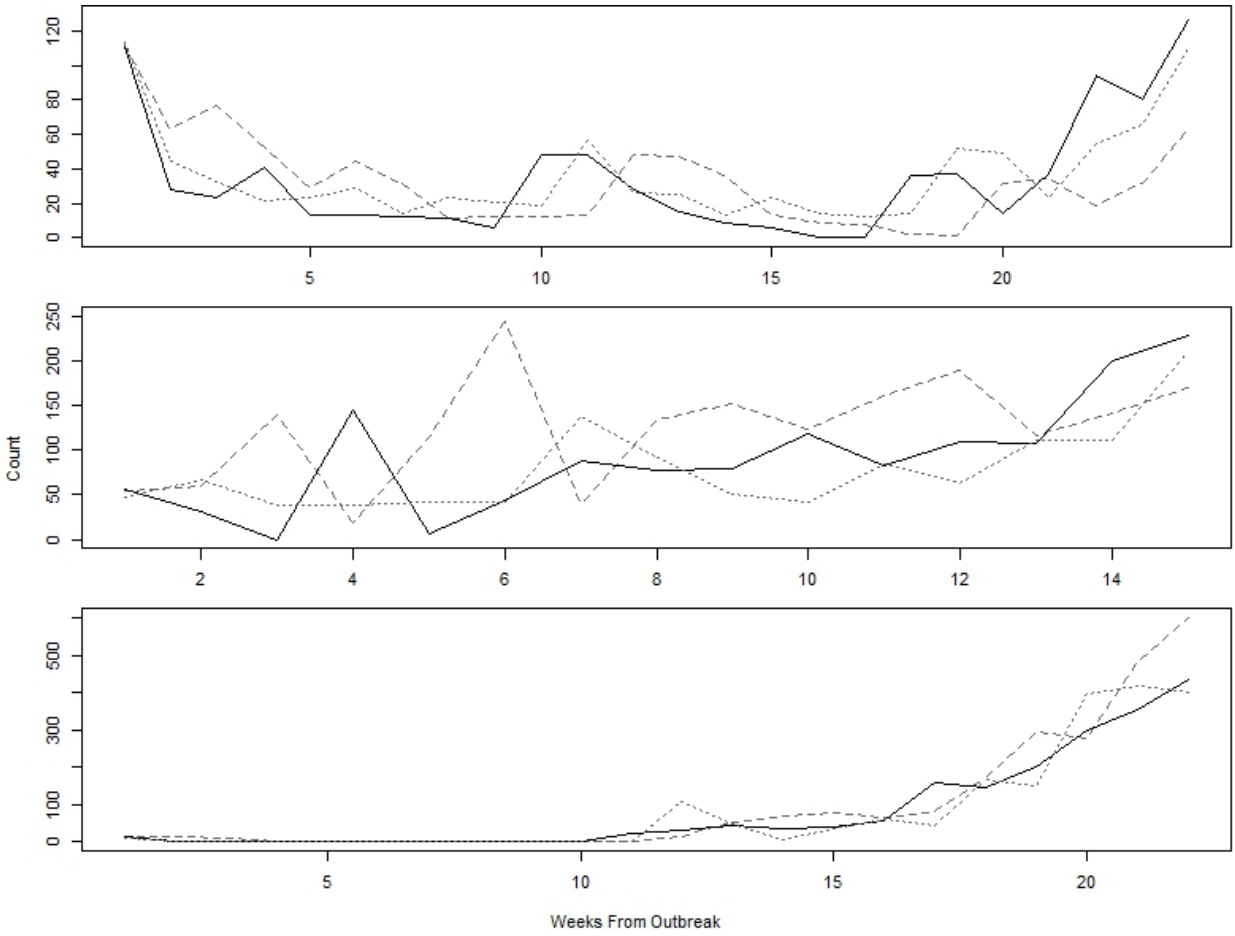
| | **Guinea** | **Sierra Leone** | **Liberia** |
|---|---|---|---|
| Background rate $\hat{\mu}$ | 0.544 | 0.91 | 0.037 |
| | (0.053) | (0.089) | (0.015) |
| Productivity constant $\hat{K}$ | 0.893 | 0.931 | 0.997 |
| | (0.011) | (0.0067) | (0.0012) |

Standard errors in parentheses are calculated following Fox et al. (2016).

Table 2: Log-likelihood, AIC and weekly RMSE for SEIR and Hawkes

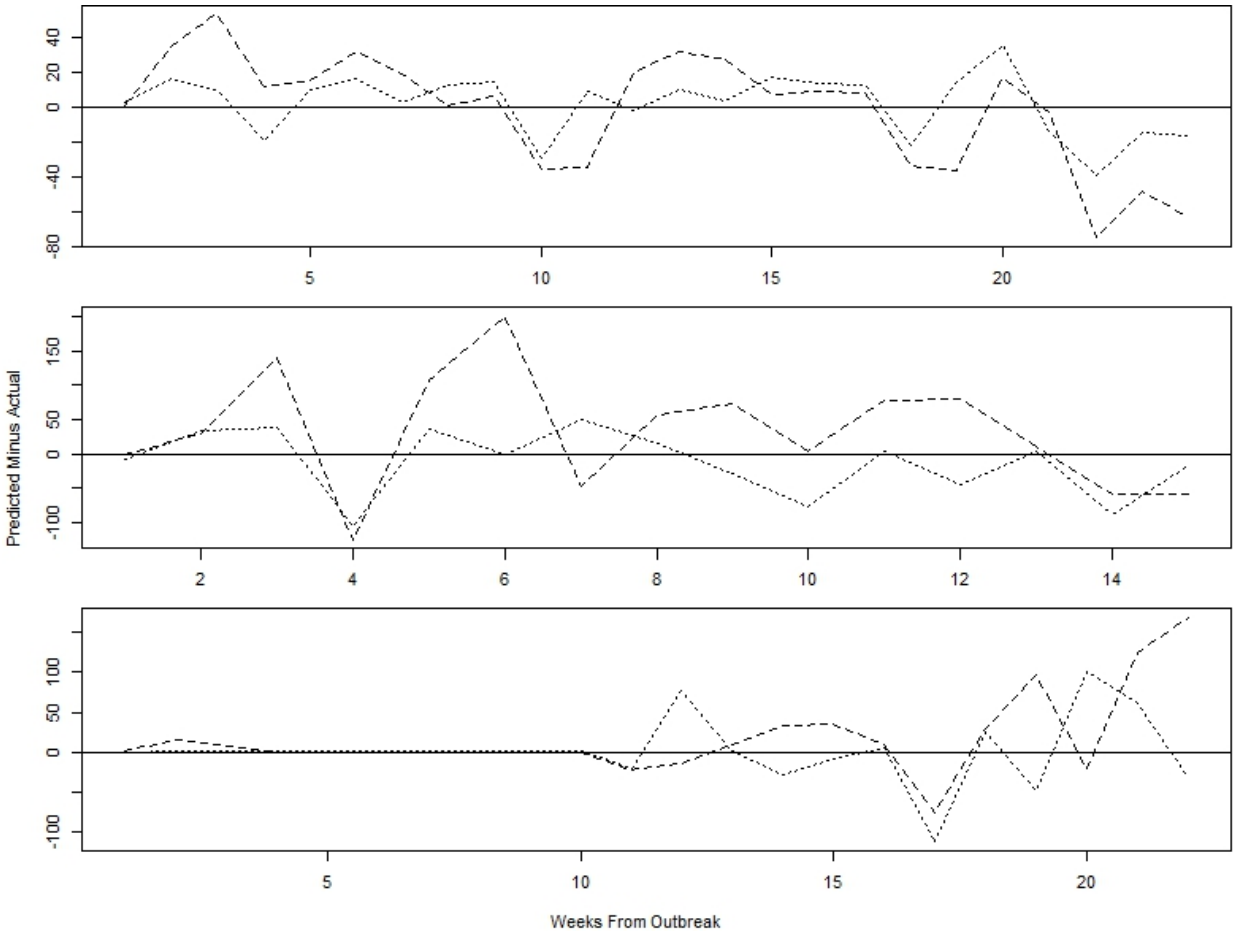
| | **Guinea (861 cases)** | | **Sierra Leone (1424 cases)** | | **Liberia (2081 cases)** | |
|---|---|---|---|---|---|---|
| | SEIR | Hawkes | SEIR | Hawkes | SEIR | Hawkes |
| Log-Likelihood | -606.5 | 919.0 | -239.3 | 2892.6 | -330.0 | 5265.5 |
| AIC | 1207.0 | -1234.0 | 472.6 | -5181.3 | 654.0 | -9933.0 |
| RMSE of weekly forecasts (cases) | 33.25 | 17.85 | 91.84 | 49.91 | 55.03 | 42.31 |

Figure 3: Weekly forecasts of new infections from SEIR and Hawkes models.



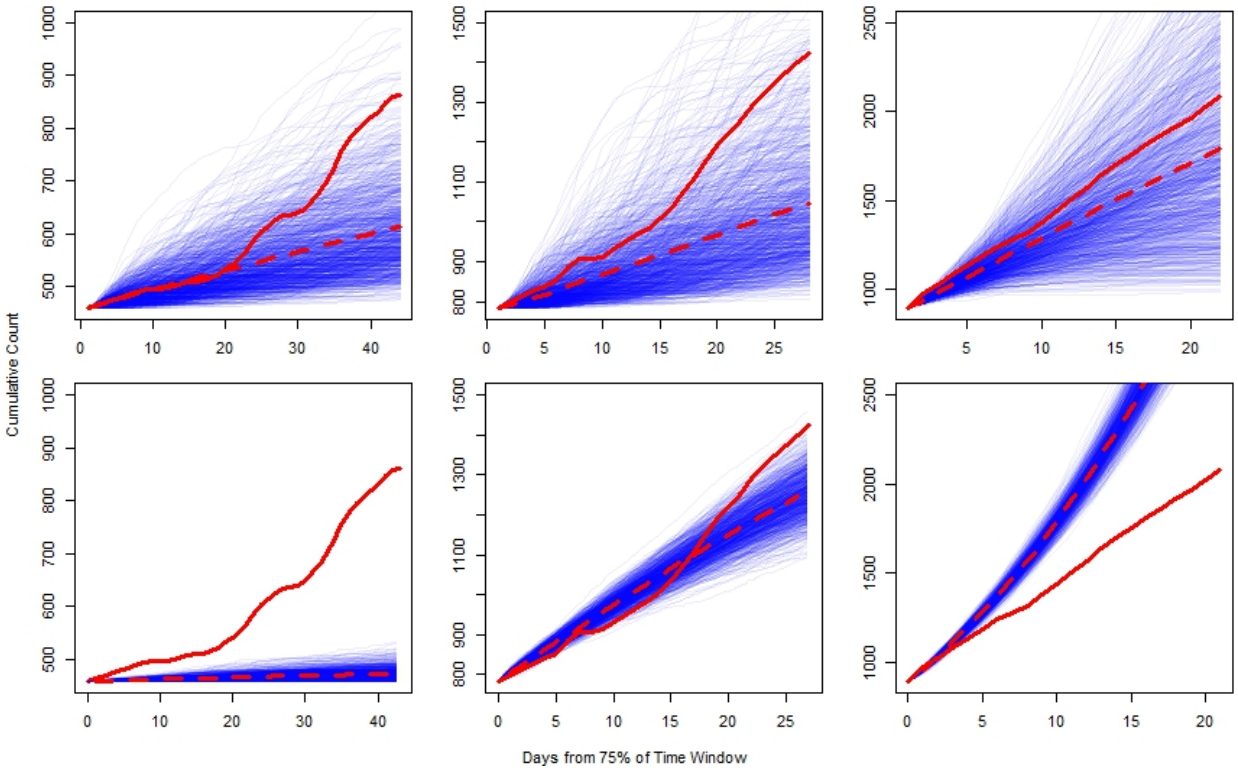

(Top to bottom): Guinea SE, Sierra Leone East, Liberia NW. Solid curve = observed new case incidence per week as reported in WHO data, dashed curve = SEIR forecast, dotted curve = Hawkes forecast. The start dates of outbreak from top to bottom are 2014-03-23, 2014-05-27 and 2014-04-05 respectively. Each weekly forecast is the mean of 1000 simulations. For each week, simulations of new cases were conducted using model parameters fitted over each country's entire data set. Each week's simulations began with the same number of initial infected cases based on the history of reported infections preceding each week's simulation start date.

Figure 4: Errors in mean of weekly forecasts from SEIR and Hawkes models.



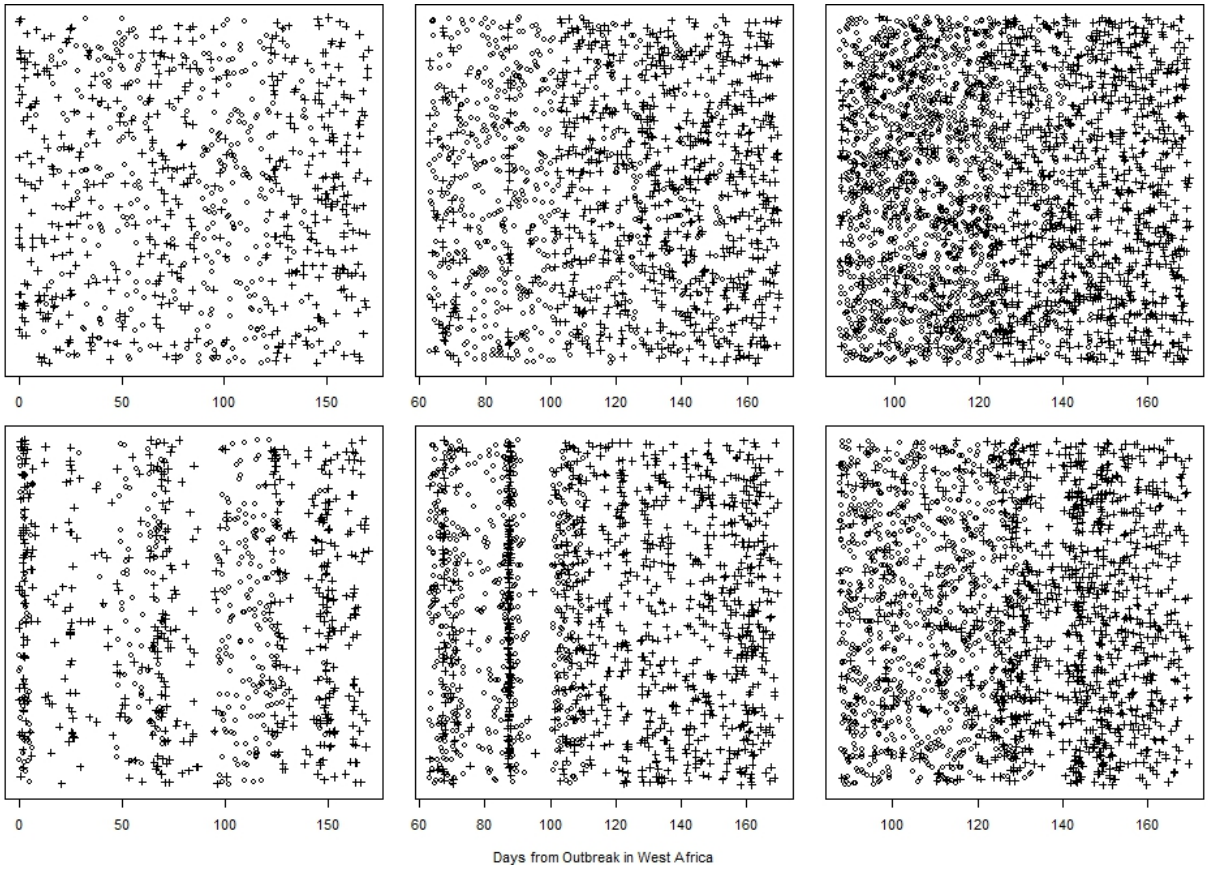

(Top to bottom): Guinea SE, Sierra Leone East, Liberia NW. Dashed curve = weekly error of SEIR model forecasts, and dotted curve = weekly error of Hawkes model forecasts. The start dates of outbreak from top to bottom are 2014-03-23, 2014-05-27 and 2014-04-05 respectively. Each weekly forecast is the mean of 1000 simulations. For each week, simulations of new cases were conducted using model parameters fitted over each country's entire data set. Each week's simulations began with the same number of initial infected cases based on the history of reported infections preceding each week's simulation start date.

Figure 5: SEIR and Hawkes projections using first 75% of data for fitting




(Left to right): Guinea SE, Sierra Leone East, Liberia NW. Starts dates of simulations from left to right are 2014-07-28, 2014-08-13 and 2014-08-19 respectively. Thin curves in top panels show 1,000 simulations of Hawkes model (2) with parameters fit using first 75% of the data for the corresponding country and simulated forward for the last 25% of the observed time period. Thin curves in bottom panels show 1,000 simulations of SEIR model with parameters fit using first 75% of the data for the corresponding country and simulated forward on the last 25%. Dashed curve = mean of simulations. Solid curve = actual cumulative total number of observed cases as reported by WHO.

Figure 6: Superthinning using Hawkes and SEIR infection rate parameters




(Left to right): Guinea SE, Sierra Leone East, Liberia NW. Top = Hawkes, bottom = SEIR. Thinned original points are marked with plus signs, and superposed points are marked with circles. X-axis indicates days from 2014-03-23, the beginning of the West African Ebola outbreak. The y-coordinates are uniform(0,1) random variables.