Title
Parametric estimation of spatial-temporal Hawkes models for the spread of Ebola in West Africa in 2014

Author
Krebs, Alex Joshua

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Parametric estimation of spatial-temporal Hawkes models for the spread of Ebola in West Africa in 2014

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Alex Joshua Krebs

2017

ABSTRACT OF THE THESIS

Parametric estimation of spatial-temporal Hawkes models

for the spread of Ebola in West Africa in 2014

by

Alex Joshua Krebs

Master of Science in Statistics

University of California, Los Angeles, 2017

Professor Frederic R. Paik Schoenberg, Chair

Parametric Hawkes models are proposed and fit by maximum likelihood to World Health Organization data from the 2014 Ebola epidemic in West Africa. Models were fit to various sub-region-level subsets of the data to compare with previous research on compartmental models and nonparametrically estimated Hawkes processes. Models were also fit to country-level subsets and multi-country subsets to evaluate how these models perform on increasing scales. Results suggest that these spatio-temporal models are able to accurately forecast the spread of Ebola infections on larger space-time windows than have been previously researched, with the benefit of improved parameter interpretability.

The thesis of Alex Joshua Krebs is approved.

Nicolas Christou

Ying Nian Wu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles,

2017

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to extend my love and thanks to the following:

To my adviser, Rick, and peers Adam, Junhyung, and Ryan who have developed and pursued this project, I am so grateful to have had this opportunity to work with you.

To my entire committee and those who helped with this project, including Rick, Rob, Nicolas, and Ying Nian, thank you for all the classes I have had the pleasure of taking under your direction, and for shaping my time at UCLA in the most positive ways.

To my counselor, Glenda, thank you for four amazing years— your help and friendship throughout has meant the world to me.

To my parents Kathy and David, my sister Katie, my grandparents Gram and Saba, and Debbie, Ned, Danny, Mike, Jeff, and Fritz, and my entire Widdop family, thank you for your unwavering support and encouragement. You have always been my primary source of inspiration.

Finally, to my one and only KG, thank you for keeping me laughing and smiling throughout this whole process. I can't wait to celebrate our upcoming successes— together!

Introduction

West Africa experienced the worst Ebola epidemic in recorded history between 2014 and 2016. Though the disease spread to seven countries in the region, the outbreak proved particularly devastating in Guinea, Sierra Leone, and Liberia, infecting over 28,000 individuals and killing more than 11,000. Despite the grim realities surrounding this virus, more than 10,000 people infected with Ebola have been cured, due in large part to heightened efforts to vaccinate, detect, and contain at-risk populations (World Health Organization, 2016).

Existing epidemiological methods for modelling the spread of disease are founded on compartmental models introduced by Kermack and McKendrick (1927). Such models, especially the Susceptible-Exposed-Infected-Recovered (SEIR) model, was proposed to describe African Ebola outbreaks by Chowell et al. (2004) and fit to the 2014 West African Ebola outbreak in Guinea, Sierra Leone, and Liberia by Althaus (2014). However, as noted in Chaffee et al. (2017), SEIR models may be overly simplistic and have large errors in practical application given their limited strictly-temporal scope and assumption that all susceptible individuals have equal probability of infection. As an alternative, Chaffee et al. (2017) propose purely temporal Hawkes models, estimated non-parametrically and individually for each spatial location. The performance of these Hawkes models in forecasting the purely-temporal spread of Ebola motivates further exploration and discussion of these methods as viable alternatives to existing SEIR models. However, given the economic and social consequences Ebola has had even in neighboring countries untouched by the virus, it is now important to consider spatial distribution and progression of Ebola (United Nations Development Programme, 2015).

Here, we propose parametric spatial-temporal versions of the Hawkes process, with parameters estimated by maximum likelihood estimation. The models are assessed using likelihood criteria, superthinned residuals, and an out-of-sample validation in which only a portion of the data is used in fitting models and the remainder is used for evaluation against multiple simulations.

Design

Outbreak Data

The data consists of WHO records from three West African countries during the 2014 Ebola outbreak (WHO, 2016). These records contain information about the country, geographic location within the country (either a region, city or village), and the number of infections and deaths associated with the virus on the date of observation. Dates of infection range from March 23, 2014 through September 7, 2014, and are typically recorded at weekly intervals. The initial subsets of the data include three sub-regions including Southeast Guinea, Eastern Sierra Leone, and Northwest Liberia, each of which is handled separately and fit with unique models.

Hawkes processes require that no two observations occur at the same point in time and space, so the dates of each batch of observations were uniformly distributed between the recorded date of observation and the date of the previous batch in a given region. Further, as it is infeasible that all infections occurred at the exact same location, observations were jittered uniformly in a circle encompassing the region in focus.

Parametric Hawkes Modeling and Evaluation

Parametrically specified spatial-temporal Hawkes models may offer several potential advantages compared to the SEIR models used by Althaus (2014) and the non-parametric Hawkes models used by Chaffee et al. (2017). First, compared to the previous studies which fit purely temporal models to each spatial region individually, additional precision may be obtained by incorporating spatial-temporal information from each observation into the model. Such precision may aid in the realism of the model as well, since spatial proximity would clearly influence the transmission of a contagious disease like Ebola. Second, the parameters in a parametric Hawkes model may offer additional interpretability and insight into the spread of the virus.

Any analytical spatial-temporal point process is characterized uniquely by its associated conditional rate process λ (Fishman et al., 1976). A separable spatial-temporal Hawkes process has conditional intensity defined by λ(t,x,y), which may be thought of as the frequency with which events are expected to occur around a particular location (t,x,y) in space-time, conditional on the prior history, H_t, of the point process up to time t (Schoenberg et al., 2013). The parametric model defines the conditional intensity λ at every point in the space-time window, and takes the form:

$$\lambda(t,x,y) = \mu + \kappa \sum_{t>t_i} g_t(t-t_i) g_{x,y}(x-x_i, y-y_i) \qquad (1)$$

We assume separability of the spatial and temporal triggering here for convenience. Our investigations of non-parametric models explored by Chaffee et al. (2017) suggest the forms for the temporal and spatial triggering functions, $g_t(t)$ and $g_{x,y}$ respectively, to be:

$$g_t(t) = \beta e^{-\beta t} \tag{2}$$

$$g_{x,y} = \frac{\alpha}{\pi} e^{-\alpha r^2} \tag{3}$$

$$r^2 = x^2 + y^2$$

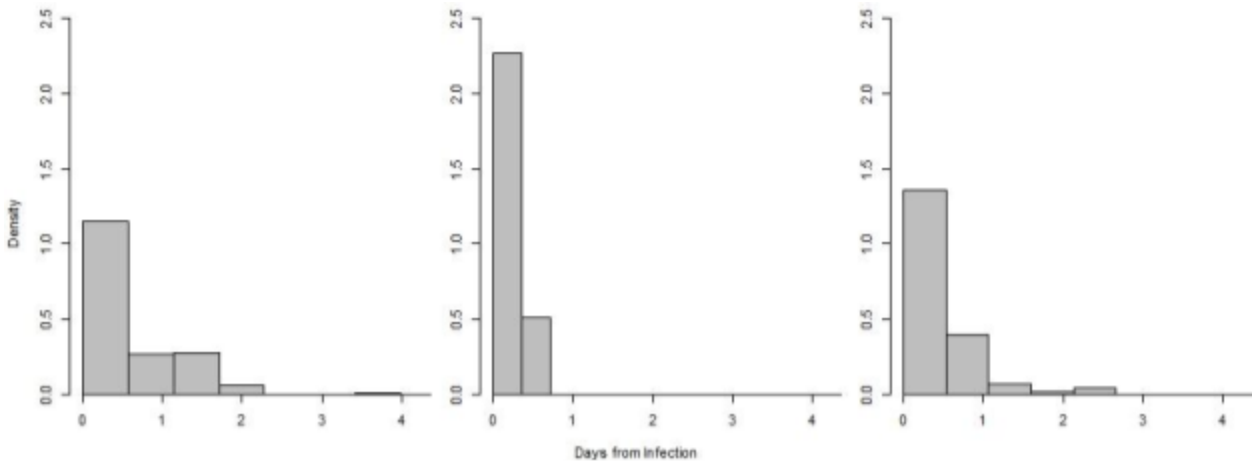

Figure 1: Estimated Hawkes triggering density for Southeast Guinea, East Sierra Leone, and Northwest Liberia (Chaffee et al. 2017)

In Formula 1, μ is the background rate and κ is the productivity or the expected number of points spawning from another. κ is most closely related to the reproductive number, R0, in SEIR models (Korobeinikov, 2004). If κ is larger than 1, the process is considered explosive and gives rise to rapid, unlimited spread. To prevent unrealistically modeling an unconstrained spread of Ebola, κ is bound between 0 and 1 such that an infected individual cannot directly infect more than one other.

In the spatio-temporal case, the log-likelihood of this process is:

$$\sum log(\lambda(t_i)) - \int \lambda(t,x,y)dtdxdy$$

The parameters are estimated on the entire data available for each sub-region using maximum likelihood, and the resulting optimized models can be evaluated by their superthinned residuals (Clements et al., 2013). Consistent with previous research, we first thin or remove each point with probability $1 - \min\{\frac{c}{\widehat{\lambda}(t_i,x_i,y_i)}, 1\}$, where c is chosen using the default suggested by Clements et al. (2013) as the mean of $\widehat{\lambda}$ over all observed points. Next, we simulate a homogenous Poisson process with rate c and keep each resulting point with probability $\max\{\frac{c-\widehat{\lambda}(t_j,x_j,y_j)}{c}, 0\}$. The resulting superthinned residuals resemble a Poisson process with rate c if and only if the estimated conditional intensity $\widehat{\lambda}$ is correct (Clements et al. 2013). The superthinned residuals are thus inspected for uniformity and can be tested using standard methods or by eye. Clustering in the superthinned residuals corresponds to areas where the model overestimated the conditional intensity, and sparsity of points in the residuals corresponds to areas where the conditional intensity was underestimated. Inhomogeneity in the forms of clusters and voids can be expected for a simple Hawkes model because of the fixed background rate μ and productivity κ, though ideally these plots will not exhibit discernible departures from uniformity.

To guard against overfitting, the records from each sub-region are modeled with the same parameters and restrictions, using the first 75% of data to predict the remaining 25%, then again using the first 50% the data to predict the remaining 50%. Each model is evaluated by comparing 200 simulations against the observed counts provided in the original WHO data.

These methods— modeling, superthinning, and simulating— are applied to the outbreaks in each country individually and also collectively to all records from the three West African Countries.

Results

Model Fitting and Residual Analysis

Estimated model coefficients, associated standard errors, and log-likelihood scores for all models are displayed in Table 1. In context, because κ is bound between 0 and 1 to prevent the scenario of limitless spread, we anticipate $\alpha > \beta$, which is confirmed by every model.

Maps of superthinned residuals for each subregion are depicted in Figure 4. Upon observation, the superthinned residuals over all sub-regions appear homogenous. Model 1, associated with Southeast Guinea and Figure 4a, appears to overpredict points in the center of the sub-region from days 130-140, which is seen as a void during this time. Ultimately, the 10-day interval for these plots is an arbitrary selection, and as we can see in days 140-150, the model seems to resume explaining activity in the center of Southeast Guinea reasonably well.

Forecast Analysis

In addition to superthinned residuals, we consider forecasting via simulation as a means of model evaluation. To compare with Chaffee et al. (2017), one subregion from each country is first isolated, and a model is fit to 50% and 75% of outbreak time ($\tau_{.50}$ and $\tau_{.75}$ respectively). For example, Southeast Guinea's observations span 169.93 days, so for the 50% time subset, the model was fit to all observations occurring before $\tau_{.50} = 169.93*0.50 = 84.96$ days. The resulting simulations are plotted alongside the observed infection counts from each subregion in Figures 5 and 6.

The Hawkes models applied to Southeast Guinea and East Sierra Leone forecast the end of the outbreak in both regions impressively well. In Liberia, a small number of isolated infections were reported at the onset (week 1), followed by numerous weeks with very few

additional reports. The estimated model coefficients in Model 9 imply the lowest productivity and rate of infection among all models and subregions. As a result, the Hawkes model applied to Liberia's data restricted to $\tau_{.50}$ estimates strikingly low counts of second-half contractions (Figure 6). In Figure 5, however, it is interesting to note that despite most simulations estimating fewer cases than reality, a few forecasts following $\tau_{.75}$ in Liberia capture the potential for an event of rapid spread.

Analyses of All Available Data

To showcase the advantage of Hawkes models— namely their consideration of space— the final component of this paper is the evaluation of all three countries in a comprehensive model. The combined data from Guinea, Sierra Leone, and Liberia is restricted and modeled with the same procedures as before (Models 10,11,12). The superthinned residual plots in Figure 7 appear impressively uniform through the first half of the outbreak, suggesting Model 10 has characterized clustering in the data appropriately, but after day 110 the residuals appear to cluster more readily. We suspect that the coarse spatial resolution of the original data is the primary driver of this behavior. Figure 8 shows narrow simulation ranges resulting from Model 11, accurately forecasting the number of new infections during the final weeks of the 2014 Ebola outbreak. Figure 9 suggests that Model 12 also forecasts infections after $\tau_{.50}$ reasonably well. These results further support a large-scale approach to model-fitting, especially since the three countries share borders and realistically should not be considered independent during an epidemic.

Discussion

Chaffee et al. (2017) found non-parametric Hawkes models to perform at least as well as compartmental SEIR models for predicting the spread of Ebola during the 2014 epidemic. Here we have shown that parametric Hawkes models also have the capacity to accurately predict the spread of Ebola with the added benefit of interpretability of estimated coefficients, which was previously unaccounted for by their non-parametric counterparts. Furthermore, even simplistic models such as the ones presented here have an enormous advantage in their ability to handle spatio-temporal data compared to the existing strictly-temporal SEIR models. However, SEIR models might still have an advantage in the early stages of an outbreak when few infections are recorded.

At present, the primary limitation of these results is the spatial and temporal resolution of the original data. The location of an infection was only recorded as the sub-region in which the disease was contracted. Though these sub-regions provide a finer resolution than the broader country, they are still vast and limited in their accuracy. The time associated with each case is also coarse, having been recorded in batches with other recent infections. With enhanced spatial and temporal resolution, Hawkes models have the potential to more accurately detail the spread of disease.

Conclusion

The performance of parametric Hawkes models on various subsets as well as the entirety of the 2014 Ebola outbreak data further validates these techniques as viable infectious disease models. Clearly not all environmental factors are taken into consideration in the present analysis,

but the capacity of simple models to accurately predict the spread of the virus shows promise for models of greater complexity. Future exploration of statistical models in an epidemiological context should consider marked point processes that take into account additional factors associated with space or time such as a city's population or status of vaccinations. Such models are already being implemented in numerous settings, including earthquake, wildfire, and epidemic occurrences— among many others (Ripley, 1997; Guttorp, 1995; Schoenberg et al. 2002). Another practical interest is the extent to which a limited amount of data can be modeled and make accurate predictions, potentially leading to optimal selections of τ for simulating future spread.

Tables

Table 1: Model Parameters

Country	Region	Subset (as percentage of time) used for Modeling	Model Number	Log-likelihood	μ	κ	α	β
Guinea	Southeast	Full	1	1662.37	0.170 (0.048)	0.966 (0.034)	136.145 (13.066)	0.397 (0.031)
		75%	2	644.32	0.205 (0.063)	0.943 (0.047)	99.276 (12.552)	0.431 (0.046)
		50%	3	604.42	0.212 (0.085)	0.955 (0.051)	109.997 (15.805)	0.416 (0.045)
Sierra Leone	East	Full	4	3550.24	0.511 (0.114)	0.961 (0.027)	149.041 (12.459)	0.657 (0.048)
		75%	5	1556.78	3.723 (0.611)	0.933 (0.048)	231.353 (30.075)	0.359 (0.034)
		50%	6	793.30	5.944 (0.770)	0.999 (0.000)	67.63 (0.132)	5.421 (1.357)
Liberia	Northwest	Full	7	6340.10	0.166 (0.046)	0.988 (0.022)	152.047 (11.900)	0.440 (0.028)
		75%	8	604.99	1.526 (0.341)	0.999 (0.000)	32.561 (3.606)	0.844 (0.112)
		50%	9	4.953	0.073 (0.034)	0.862 (0.146)	20.032 (6.062)	1.564 (0.316)
All Countries		Full	10	29848.23	0.069 (0.024)	0.983 (0.014)	196.574 (3.660)	1.518 (0.057)
		75%	11	11261.06	0.273 (0.053)	0.983 (0.022)	1590.740 (80.726)	1.107 (0.050)
		50%	12	4848.08	0.248 (0.060)	0.936 (0.030)	1679.180 (131.990)	0.654 (0.042)

Figures

Figure 2: Map of West Africa, with Guinea, Sierra Leone, and Liberia highlighted.

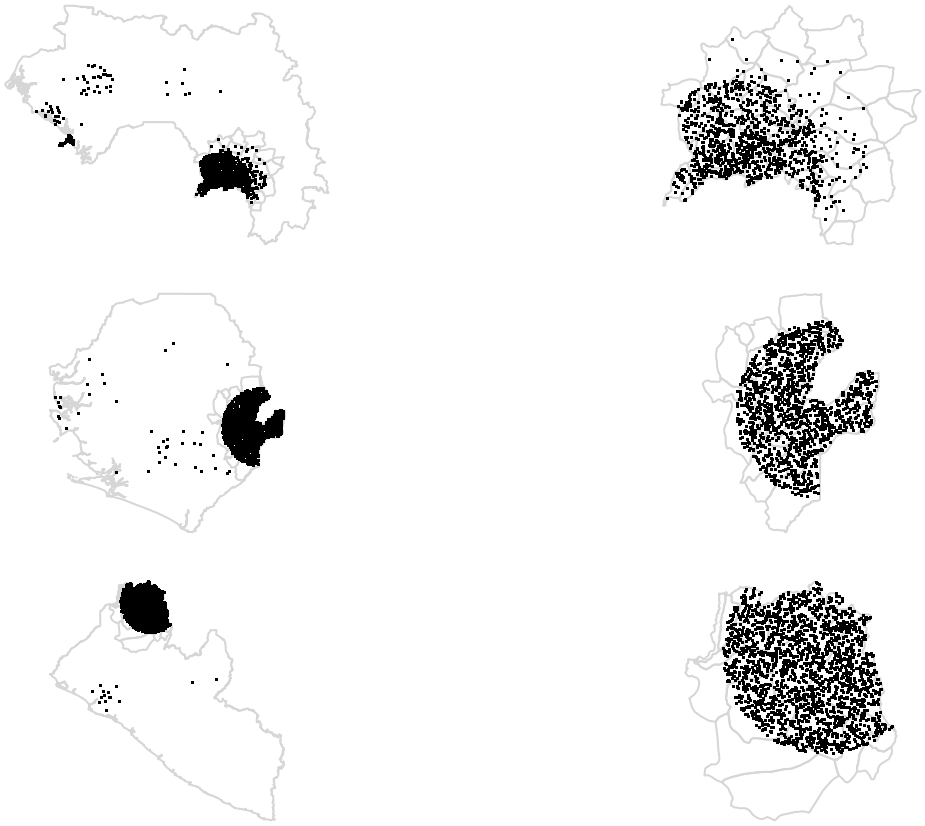

Figure 3: From top to bottom, Guinea and Southeast Guinea, Sierra Leone and East Sierra Leone, and Liberia and Northwest Liberia with 2014 WHO observations shown.

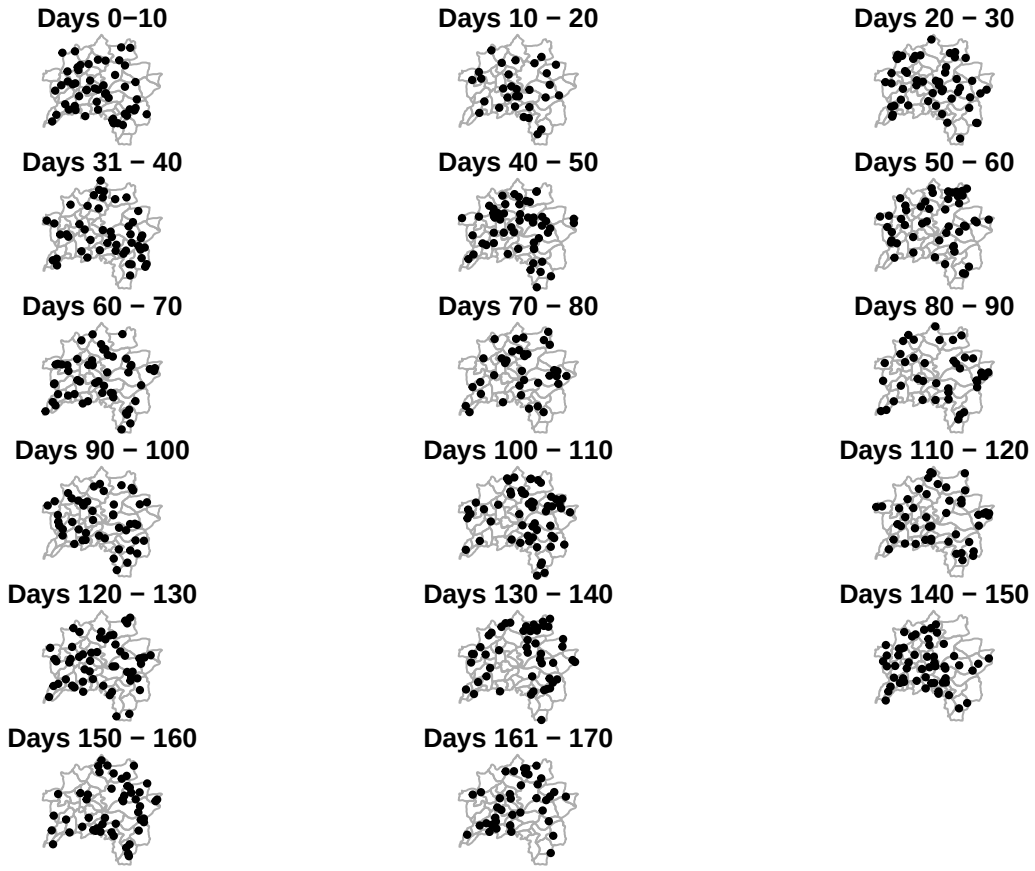

Figure 4a: Superthinned residuals resulting from Model 1 plotted over Southeast Guinea in 10-day intervals.

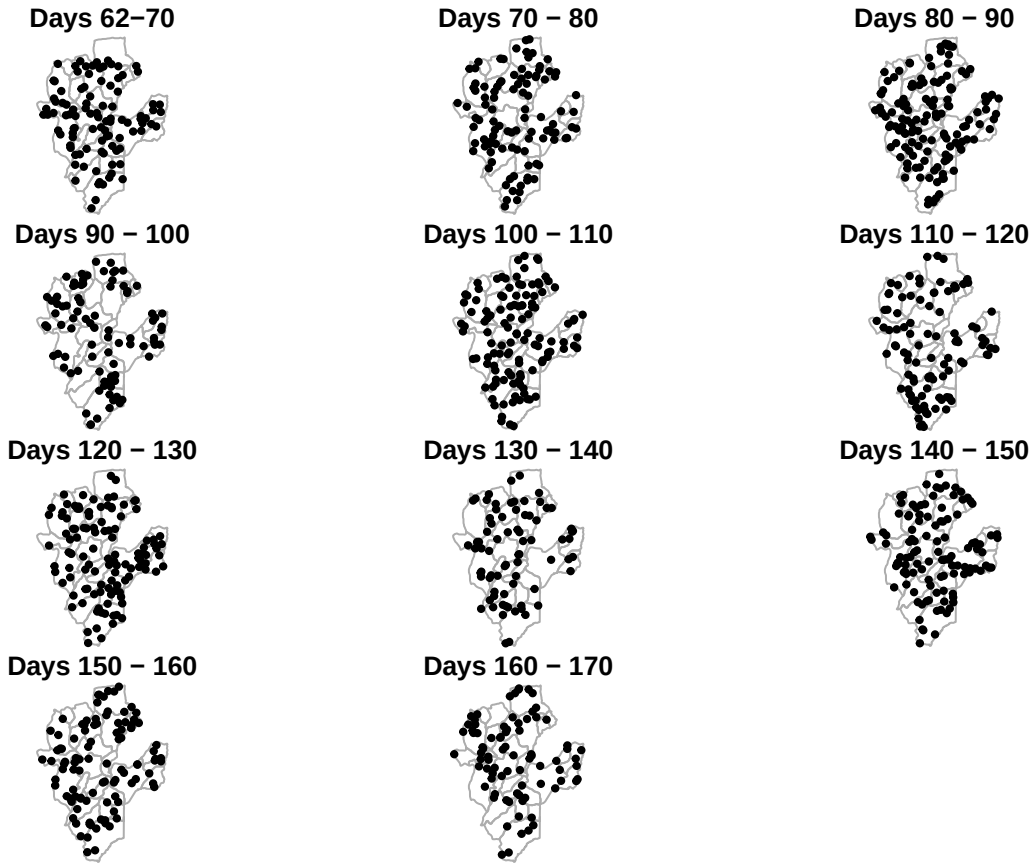

Figure 4b: Superthinned residuals resulting from Model 4 plotted over East Sierra Leone in 10-day intervals.

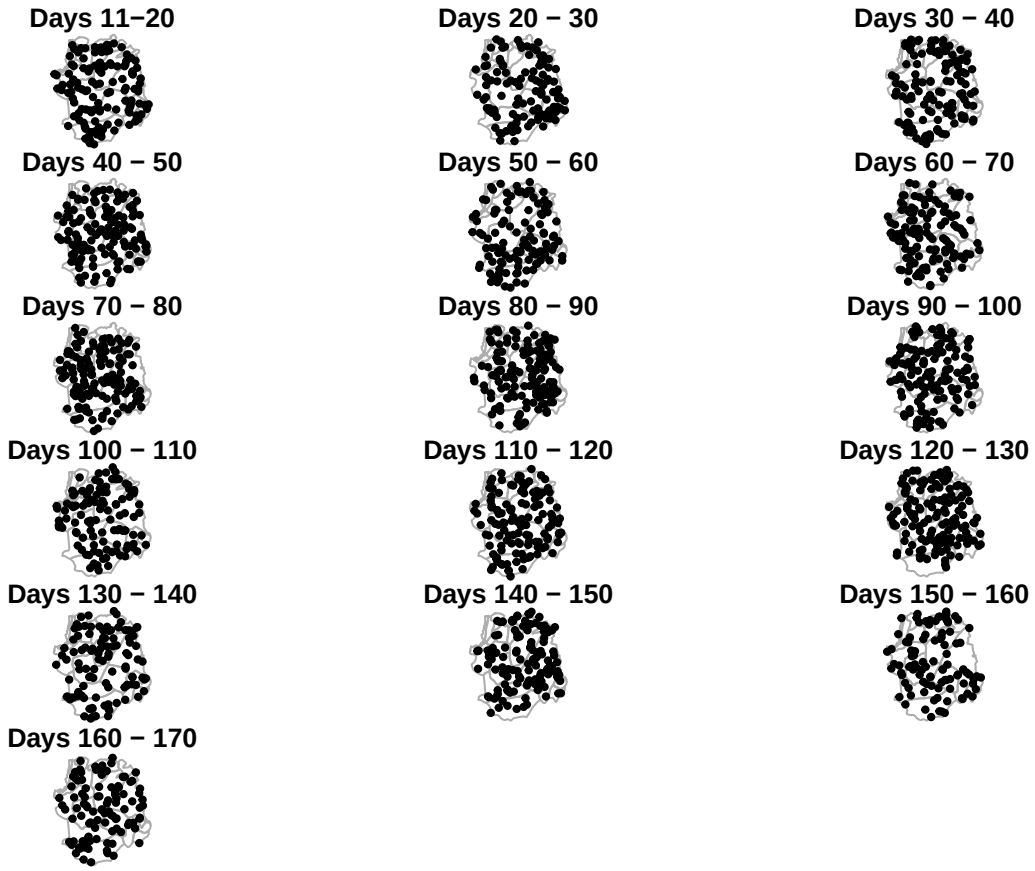

Figure 4c: Superthinned residuals resulting from Model 7 plotted over Northwest Liberia in 10-day intervals.

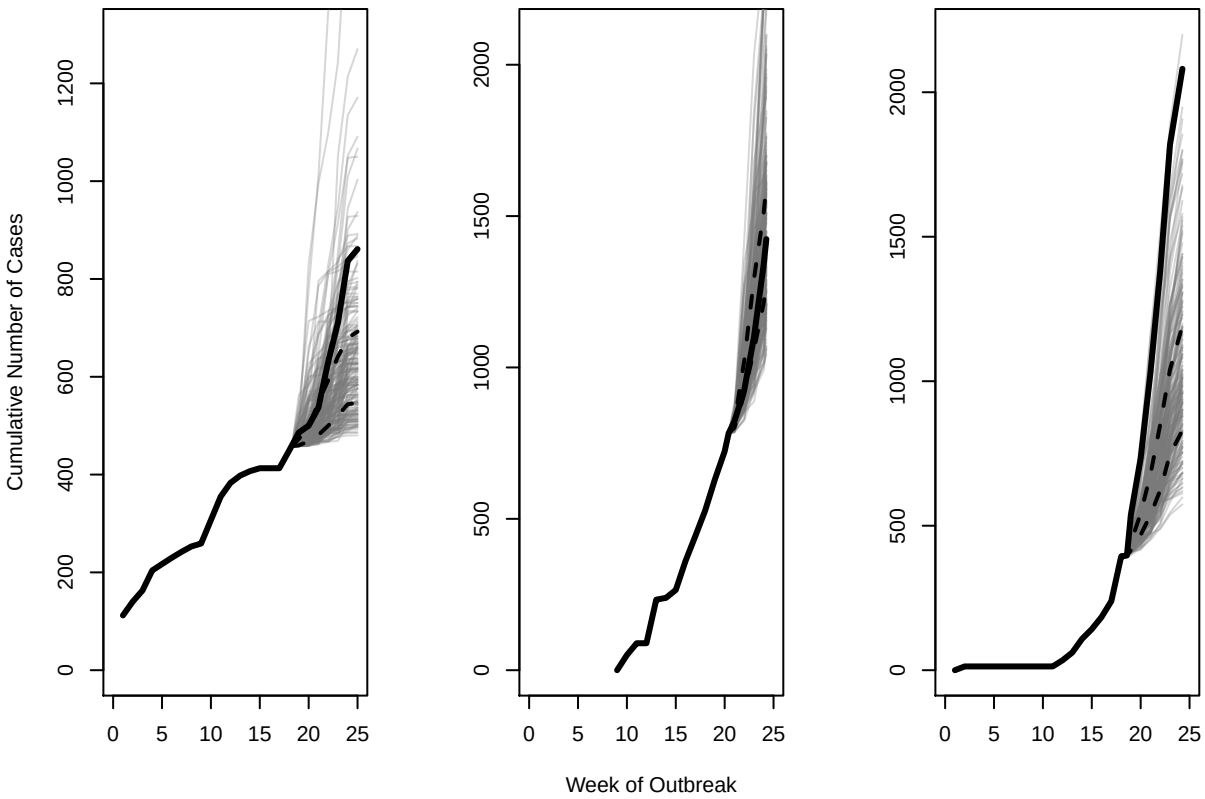

Figure 5: From left to right, simulations of the final 25% of outbreak in Southeast Guinea, East Sierra Leone, and Northwest Liberia. The solid black curve is the true data, the slightly transparent grey curves represent unique simulations, and the dashed lines depict the 25th and 75th percentiles of all simulations for reference. $\tau_{.75}$ are 127.45, 143.14, and 130.21 days respectively.

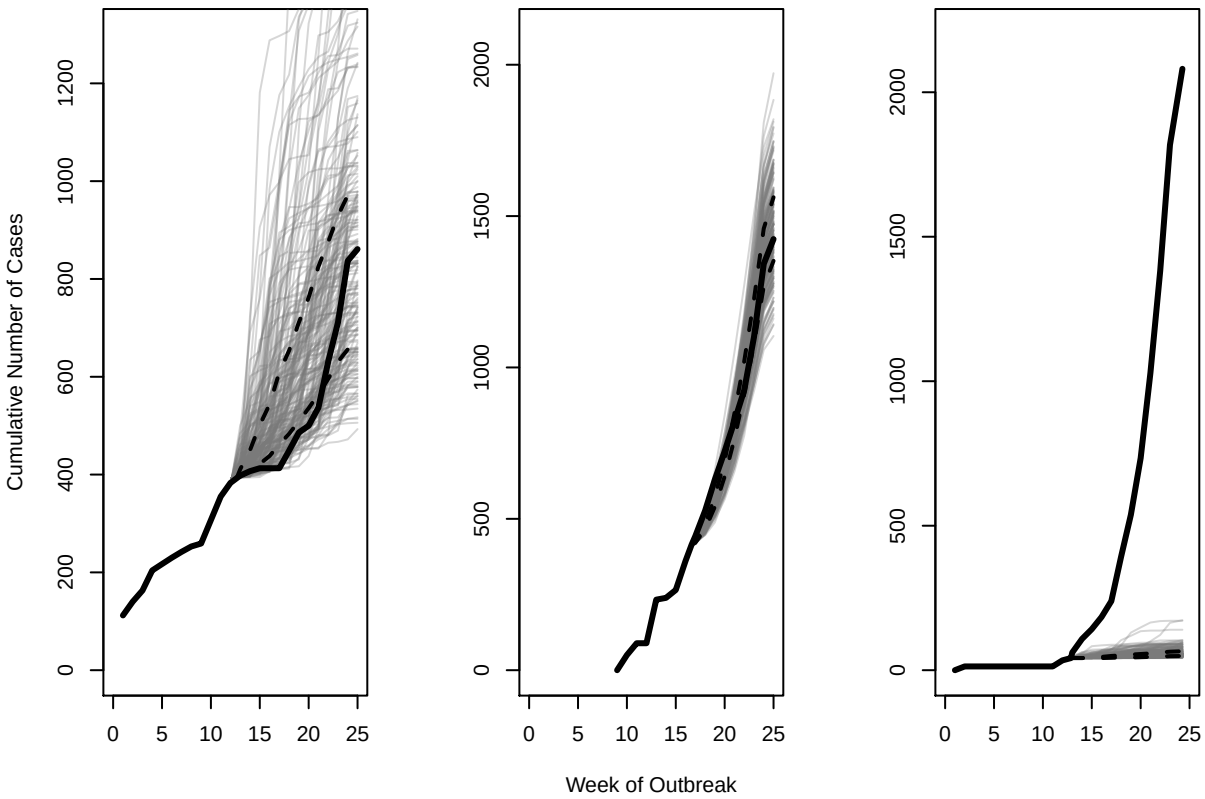

Figure 6: From left to right, simulations of the final 50% of outbreak in Southeast Guinea, East Sierra Leone, and Northwest Liberia. The solid black curve is the true data, the slightly transparent grey curves represent unique simulations, and the dashed lines depict the 25th and 75th percentiles of all simulations for reference. $\tau_{.50}$ are 84.96, 116.41, and 90.52 days respectively.

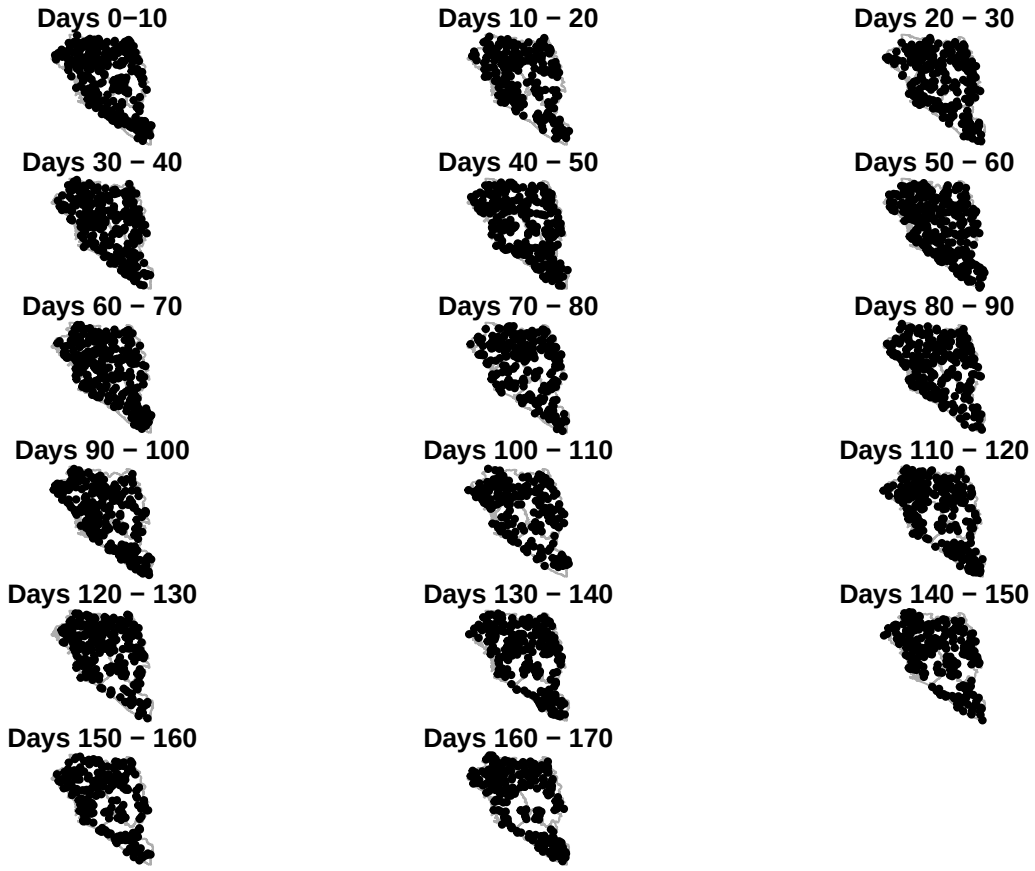

Figure 7: Superthinned residuals resulting from Model 10 plotted over Guinea, Sierra Leone, and Liberia collectively.

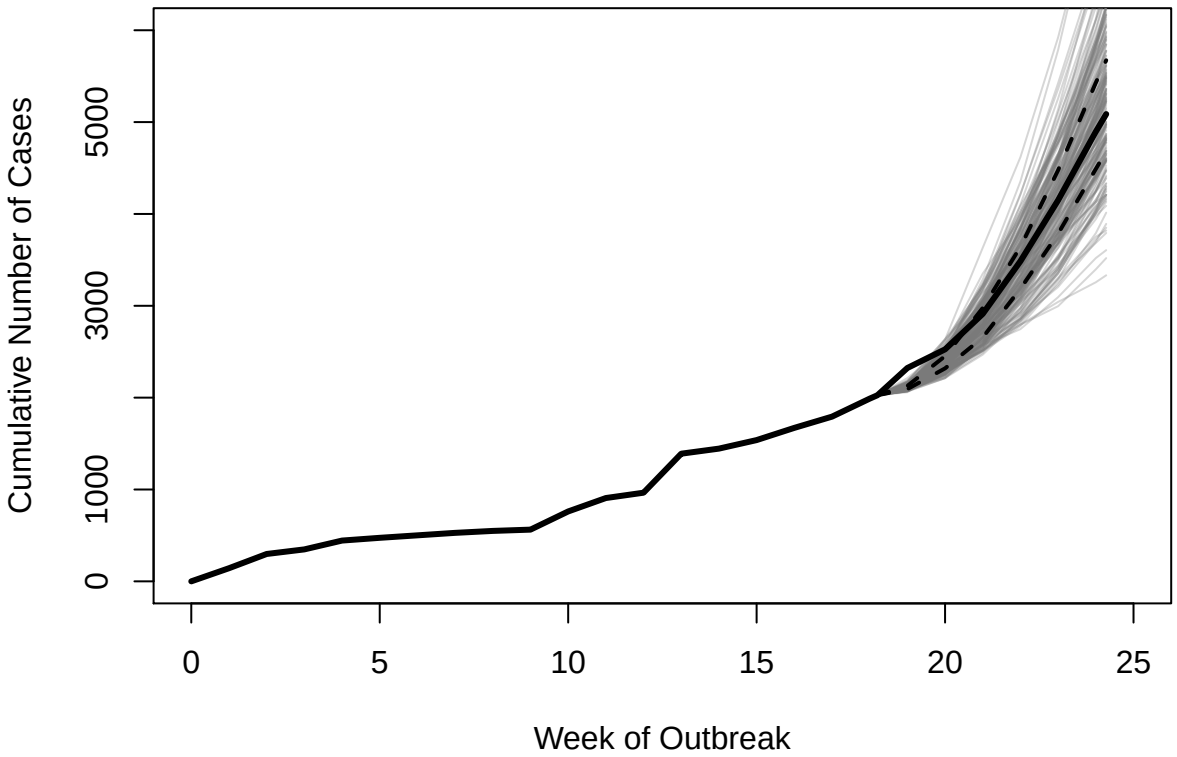

Figure 8: Simulations of the final 25% of outbreak across Guinea, Sierra Leone, and Liberia collectively. The solid black curve is the true data, the slightly transparent grey curves represent unique simulations, and the dashed lines depict the 25th and 75th percentiles of all simulations for reference. $\tau_{.75} = 127.45$ days.

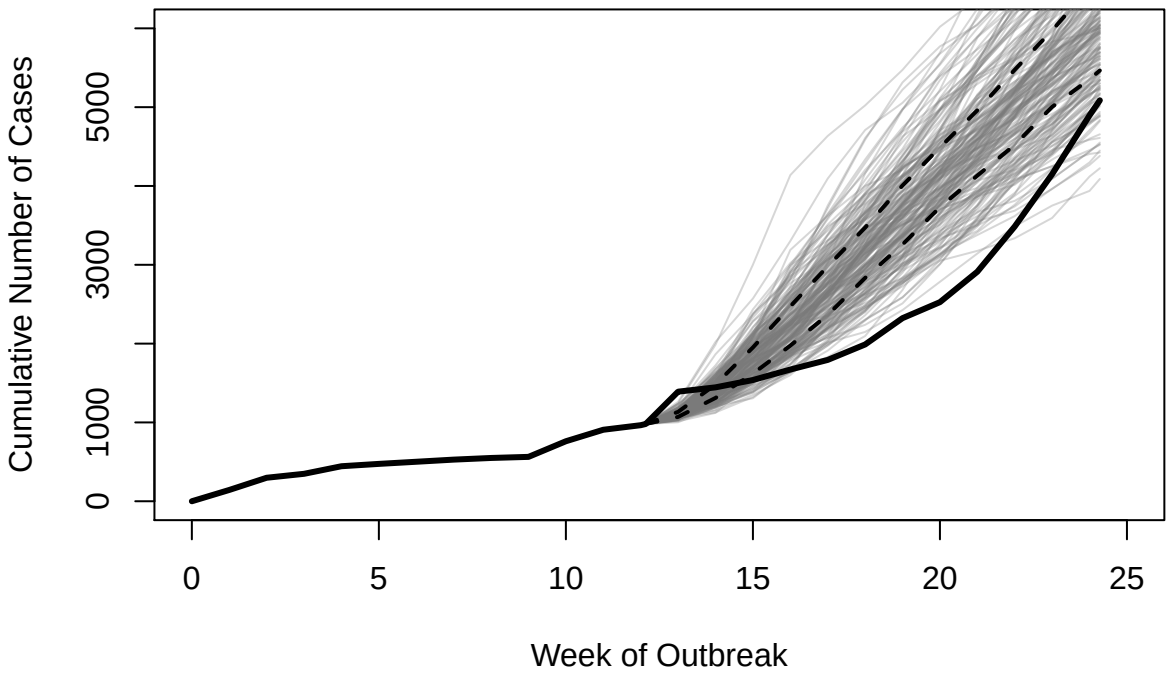

Figure 9: Simulations of the final 50% of outbreak across Guinea, Sierra Leone, and Liberia collectively. The solid black curve is the true data, the slightly transparent grey curves represent unique simulations, and the dashed lines depict the 25th and 75th percentiles of all simulations for reference. $\tau_{.50} = 84.96$ days.

References

1. Althaus, C.L. (2014) “Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa.” In: PLOS Current Outbreaks.

2. Chaffee, A.W. (2017). Comparative analysis of SEIR and Hawkes modeling as tools to characterize the 2014 West Africa Ebola outbreak.

3. Chowell, G. et al. (2004). “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, pp. 119–126.

4. 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.

5. Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, volume I: *Elementary Theory and Methods of Probability and its Applications*, 2nd edition. New York: Springer-Verlag.

6. Fishman, P.M. & Snyder, D.L. (1976). The statistical analysis of space – time point processes, *IEEE Transactions on Information Theory* **IT-22**, 257–274.

7. Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. London : Chapman and Hall.

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

9. Hawkes, A.G. and Oakes, D.A. (1974). A cluster process representation of a self-exciting process. J. Appl. Prob. (11), 493-503.

10. 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.

11. Korobeinikov, A. (2004). Lyapunov functions and global properties for SEIR and SEIS epidemic models in MMB IMA 66: 879.

12. Penttinen, A., Stoyan, D., & Henttonen, H. M. (1992). Marked point processes in forest statistics. *Forest science*, *38*(4), 806-824.

13. Ripley, B. (1977). Modelling spatial patterns. *JRSS-B* 39, 172–192.

14. Schoenberg, F.P. (2012). Point processes, spatial–temporal. in Encyclopedia of Environmetrics, Second Edition, A.H. El-Shaarawi and W. Piegorsch (eds), John Wiley & Sons Ltd, Chichester, UK, pp. 2885-2886.

15. Schoenberg, F.P., Brillinger, D.R., and Guttorp, P.M. (2013). Point processes, spatial-temporal. in Encyclopedia of Environmetrics, Abdel El-Shaarawi and Walter Piegorsch, editors. Wiley, NY, vol. 4, pp 1573–1578.

16. 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.

17. World Health Organization (2016). Ebola data and statistics. Web. Accessed September 4, 2017.