# Real-time probabelistic forecasts of the 2018 − 2019 Ebola outbreak in the Democratic Republic of Congo

CANDIDATE NUMBER

September 2, 2019

Word count: XXXX

**Abstract**

Abstract

# Contents

# 1    Introduction

Real time modelling of infectious diseases can play a crucial role during disease outbreaks like the ongoing Ebola outbreak in North Kivu in the Democratic Republic of Congo (DRC). During the previous large outbreak of Ebola in Western Africa in 2014–2015, many different models were developed both during the outbreak and after[1]. Models were used both to increase our understanding of the outbreak and to predict the spread of the outbreak. For example, to estimate if there would be sufficient bed capacity [2]. To effectively use results from mathematical modelling in an outbreak it is of key imprtance that modelling outputs are integrated into the outbreak response and decision making [3]. To acheive this requires improved communication and embedding of modelling and modellers into the outbreak response.

For forecasts to be useful in an outbreak situation, the models need to incorporate uncertainty into the predictions. Without understanding the range of possible outcomes from a model and their associated probabilities it is very difficult to support effective public health action in outbreak situations. Evaluting such probabelstic forecasts requires us to not just assess the accuracy of the point predictions, but to assess if the model correctly assess it's own uncertainty. If a model can do that, we say that it is calibrated. We are also interested in the sharpness and bias of the models, to assess their performance. These measure tell us about the absolute performance of the model, but to accurately compare models it is often better to use proper scoring rules.

In this thesis, we will describe a framework for flexibly producing probabelistic forecasts and assess how well the resulting models can be used in the onogin Ebola outbreak in North Kivu. We will compare different models to choose the best model and to explore the epidemiology of the current outbreak.

# 2    Background

## 2.1    Ebola outbreak in North Kivu and Ituri provinces

On the 1st of August 2018, the minstry of health in the Democratic Republic notified the WHO about a new Ebola outbreak in the North Kivu province [4]. North Kivu is a populous region in north eastern DRC that borders both Rwada and Uganda. By the 27th of September the WHO assessed that the outbreak constiuted a very high national and regional risk[5].

Due to distrust from the local community the outbreak response has been challenging. This has lead to difficulty in tracing contracts, giving vaccinations and treating patients. There have been many episodes of violence directed towards the ebola responders which has led to fatalities and multiple stops in the response work[6, 7].

During the outbreak, the experimental vaccination rVSV-ZEBOV-GP has been used. Preliminary data shows positive signs that this vaccine might provide some protection from Ebola [8]. The vaccine has been used in a ring vaccination strategy and been given to healthcare workers and children. This outbreak has also seen a randomised trial for medicines to treat ebola. The trial was terminated at the 12th of August due to findings that two of

the four drugs that were included showed clear effect of reducing mortaltiy from Ebola, especially with early treatment [9] At the time of writing the outbreak is still ongoing with new cases from 16 health zones and with fears that the disease will spread to larger cities in the area like Goma or accross borders. On the 17th of July the Director General of the WHO decleared the Ebola outbreak a Public Health Emergency of International Concern [10].

# 3    Modelling of Ebola

A large number of different mathematical models have been proposed for modelling Ebola, see e.g [1, 11]. The models can be divided into categories based on the structure of the models, mechanistic, semi-mechanistic, phenomenological and hybrid models. Mechanistic models are based on a specified mechanism for disease transmission and includes standard compartmental models while phenomenological models do not include any mechanism for disease transmission. An example of a phenomenological model would be an ARIMA model for the incidence. A semi-mechanisitc model include some mechanisitc asumptions, but not a full model for disease transmission.

The RAPIDD Ebol Forcasting challeng in 2018 [11] presented a range of different scenarios of an ebola outbreak that was similar to the west african outbreak and invited different groups to submitt forecasts and compared them. They found that for short-term forecasting was not related to model complexity and that "light" non-parametric models were able to do well. The more complex mechanistic models would on the otherhand be better suited to understand the effects of interventions.

In this thesis we will present a semi-mechanisitc model that includes some mechanisitc assumptions about serial intervals and forces of infection combined with flexible time-series methods. The aim is a flexible model with parameters that are easily interpretable.

# 4    Methods

## 4.1    Data

We will use daily incidence data from the North Kivu Ebola outbreak as reported by the Ministry of Health in Congo and distributed via the humanitarian data exchange (https://data.humdata.org/da cases-and-deaths-drc-north-kivu) before the 1st of September 2019. We will use the incidence of confimed cases at the national and at the health-zone level. Even with confirmed cases, there are a few days in some health zones where the incidence is negative as confrimed cases are removed. These days we set the incidence to 0.

## 4.2    Model

We consider a model where the daily incidence, $I_t$ follows a branching process. Similar models have been considered in [12, 13, 14] and have been used to model Ebola during the 2014-2016 Ebola outbreak in Wester Africa [15, 16]. In this class of models each infectious person gives rise to $\nu$ new infections, where $\nu$ is a random variable distributed with an

offspring distribution with an expected value given by the reproduction number $R$. We use the serial interval to model the time between each generation of infections. This gives a process where the expected number of new cases at time $t$ is given by the force of infection $\lambda_t$. The force of infection is a product of previous incidence weighted by the serial interval $w_\tau$ and the instaneous reproduction number $R_t$.

$$E(I_t) = \lambda_t = R_t \sum_{s=1}^{t-1} I_s w_{t-s}$$

To fully specify the model we need to determine the probability distribution for $I_t$, the serial interval and the reproduction number as a function of time. For the serial interval we will use a gamma-distribution with mean 15.3 days and standard deviation of 9.3 days as fitted to data from the West-Africa Ebola outbreak [15]. Our general approach will be to use a few different spesifications for the $I_t$ probability distribution and for estimating the time-varying reproduction number and then assess which model gives the best fit to the data.

### 4.2.1 Offspring distribution

The offspring distribution gives us the distribution of the number new infections from each infected person. By definition the expected value of this distribution is the reproduction number. The offspring distribution is a discrete probability distribution over the non-negative integers. The simplest such distribution is the poisson distribution, it is given by

$$P(X = k) = \exp{-\lambda} \frac{\lambda^k}{k!},$$

with mean and variance given by $\lambda$. This distribution implies that the number of new infections over a time interval is described by a constant rate where the chance of a new infection in each small sub-interval is independent.

For a number of diseases it has been shown that offspring distribution has larger variance than that given by the poisson distribution and has so called superspreaders [13]. One common way of modelling this over-dispersed distribution is by the negative binomial distribution. The binomial distribution describes a process where we have a series of indepdent chances to infect a new interval each with probabilty $p$, where the process continues until we have had $r$ infection chances without an infection. The distribution is given by

$$P(X = k) = \binom{k + r - 1}{k} (1 - p)^r p^k,$$

with $E(X) = \frac{pr}{1-p}$ and $Var(X) = \frac{pr}{(1-p)^2}$. We will use a parameterisation of the negative binomial distribution where we use the mean, $\mu$ and the dispersion parameter $k$. In this formulation the variance is given by $\mu(1+\frac{1}{k})$. For $k -> \infty$ we would get a poisson ditribution. The smaller $k$ is the more the outbreak is dominated by a few super-spreaders and the larger $k$ is the more similar the impact of each infected person is. For SARS it has been found that $k = 0.16$[13]. For the Ebola outbreak in West Africa they found $k = 0.53$ using conservative assumptions and data from the network of infections [16].

**PLOT of Distributions?**

The distribution for $I_t$ would be given by the sum of the offspring distributions for each infected person (weighted by the serial distribution). For both the poisson and negative binomial distribution the sum of independent distriubtions will give back the same distribution with an expected value given by the sum of the expected values. For the negative binomial distribution the dispersion parameter $k$ will remain the same.

We will investigate models with offspring distributions based both on the poisson distribution and the negative binomial distribution to assess which models fit the current outbreak better.

### 4.2.2   Reproduction number

The final ingredient to completely specify the dynamics of the model is the evolution of the reproduction number with time. Depending on the functional form of $R_t$, the model can fit a large range of possible epidemic behaviours.

It would be possible to specify a complete model for $R_t$ analytically or otherwise, but in this thesis we will instead estimate $R_t$ from the existing data and use these values to predict the future evolution of $R_t$. Since the focus of the current study is to provide a flexible framework for probabilistic prediction of disease outbreaks, this appraoch is suitable.

To estimate the reproduction number from the historical incidence data we use the method developed in [12]. If the reproduction number is estimated daily it is likely to varry too much from day to day in a manner that it is unrelastic. Therefore this method averages over the last 7 days to get more stable estimates. A baysian procedure is used to estimate both the best fit value and the uncertainty of the estimate. We use the R-package EpiEstim [17] to estimate the reproduction number using a parametric gamma distribution for the serial interval with a mean of 15.3 days and a standard devitaion of 9.3 days [15]

Once we have calculated the historical values of $R_t$ we can use them to provide an estimate of $R_t$ moving forward that can be used for predictions. We will use two different procedures to predict the reproduction number. The first method is a very simple method were we assume that reproduction number remains constant from the last historical value. We use the uncertainties given by the method in [12] to provide estimates of the uncertainty of this estimate.

The second approch to predict the reproduction number is based on baysian structural time series fitted to the historical values of $R$. We use model with a semi-local linear trend to allow the estimation of a trend in the recent data. To ensure that predicted values of $R$ are between 0 and 15 we fit the time series on a transfomed scale:

$$r^* = log\left(\frac{R}{15 - R}\right)$$

We use the R-package BSTS [18] to fit the basysian structural time series. The model we use is specified as follows

$$r^*_{t+1} = r^*_t + \delta_t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\mu),$$
$$\delta_{t+1} = D + \phi(\delta_t - D) + \eta_t, \eta_t \sim N(0, \sigma_\delta).$$

$\delta_t$ is the semi-local trend that we model as an AR(1) process that can oscilate around a level $D$. Inverse gamma-priors are used for the standar deviation parameters $\sigma_\mu$ and $\sigma_\delta$, a gaussian prior on $D$ and a $N(0, 0.1)$ prior for $\phi$. The $\phi$ parameter determines how much the trend behaves as a random-walk. For $\phi = 1$ the trend follows a random walk, while for $\phi = 0$ the trend is just constant with some noice. We use a prior for small values of $\phi$ to keep the model for having very rapid growth in the variance of $r*$ that would be unphysical.

A Markow Chain Monte Carlo (MCMC) algorithm is used to estimate the parameters in the model which then allows us to predict future values of $r*$ with uncertainties, we then transform back to the $R$.

### 4.2.3   Simulating from the model

Once we have specified our model by specifying the offspring distribution and the method for forecasting $R_t$ we can use the model to generate probabilistic forecasts. If we want to generate a forecast for $I_{t+1}$ we first use all the data up until time $t$ to estimte $R_t$ and potentially fit the time series model to those values. Our probabelistic forecast will be based on sampling possible outcomes to generate a distribution of outcomes. We therefore first sample a value for $R_{t+1}$ from our model, then we combin this with the historical incidence data to calculate $\lambda_{t+1}$. We then sample $I_t$ from the specified offspring distribution. If we want to forecast over multiple time-steps we follow the same procedure by sampling values for the reproduction number. When calculating $\lambda_{t+2}$ we use the sampled value for $I_{t+1}$ together with the historical data $I_s$ for $s \leq t$.

## 4.3   Assessing probabilistic forecasts

The aim of probabilistic forecasts is to predict both the correct average value and an appropriate uncertainty. Therefore it is not sufficient to only use metrics that depend on a point estimate, for example the Root Mean Square Error. We will follow the paradigm of maximizing sharpness of the predictive distribution subject to calibration [19]. In addition we will consider proper scoring rules for comparing probability ditributions. We follow the approach taken in [20], where probabilistic forecasts for the West African ebola outbreak were assessed using similar methods.

A model is calibrated if the uncertainties are accurate. For example if we predict that it will rain with 60% and we find that over time it does rain 60% of days where we predicted a 60% chance of rain, the model would be well calibrated. Mathematically, if we assume that real disribution of outcomes in nature is given by a cumulative density function $G_t$ and our model predicts a cummulative density function $F_t$, we say that the forecast is ideal and perfectly calibrated if $F_t = G_t$. To assess calibration we will use a randomised Probability Integral Transformation (PIT) [21]. We calculate

$$u_t = F_t(k_t) + \nu(F_t(k_t) - F_t(k_t - 1)),$$

where $k_t$ is the observed value at time, $t$ and $\nu$ is a standard uniform random variable. If the prediction is ideal, the $u_t$ will be distributed as a standard uniform distribution. We can then use the Anderson-Darling test of uniformity (goftest [22]) to assess if we can reject that the models are calibrated. Important to note that uniform PIT is a nescessary, but

not suffcient condition for an ideal forecast. In addition to assessing calibration, the PIT histogram can tell us if the forcast is under or overdispersed. If the forcast is too dispersed the PIT values are likely to cluster in the centre of the PIT histogram, while if the forcast is underdispersed they are likely to clusteralong the edges of th histogram. We use a simple measure of centrality, which is equal to the fraction of $u_t$ values that are between 0.25 and 0.75 as a way to assess if the forecasts are under or overdispersed if they are not calibrated. For both the test for uniformity and centrality we repeat the calculations 10 times and take averages to average out the effect of the randomness in the definition of $u_t$.

Sharpness is defined as the range of values in the forecast. The sharper a forecast, the more certain we are of predicted value. Sharpness depends only on the forecast and not on the observed values. We will use the normalised absolute devitation about the median of y to quantify sharpness:

$$S_t(F_t) = \frac{1}{0.675} \text{median}(|y - median(y)),$$

the normalisation factor means that $S_t$ is equal to the standard deviation if $F_t$ is normal.

It is also of importance to assess the bias of the forecast. Are we more likely to predict to large or too smal values? We will qunatify bias as

$$B_t(F_t, k_t) = 1 - (F_t(k_t) - F_t(k_t - 1)).$$

If $B_t = 0$ half the probability mass is above and half below the observed value, and the the forecast is unbiassed. $B_t$ is between -1 and 1, with both extreme values signifying a completely biased forecast.

Proper scoring rules have been developed to rank forecasts. They combined calibration and sharpness and give a consistent ranking of forecasts. We will use the the contineously ranked probabilty score(CRPS) and the Dawid-Sebastiani score (DSS) as implemented in the scoringRules package [23]. The CRPS score is given by

$$CRPS(F_t, k_t) = \int_R (F_t(z) - \mathbb{1}k_t \le z)^2 dz,$$

and the DSS only depens on the mean, $\mu_p$ and standard deviation, $\sigma_p$ of the predictive distribution

$$DSS(F_t, k_t) = \left(\frac{k_t - \mu_p}{\sigma_p}\right)^2 + 2\log\sigma_p.$$

The DSS allows an intuitive understanding of the proper scoring rules. The first term tells us about the calibration of the predictions and the second term gives information about the sharpness. For our models we will only have samples from the predictive distribution. To calculate the CRPS a kernel density estimate is used to estimate $F_t$, for the DSS we use the mean and standard deviation of the sample.

## 4.4 Implementation

The models were implemented in the R programming langauge [24] and is available open source at http://github.com/gulfa/msc_ebola. We will consider four differrent models to assess which model fits the data best. The models are:

1. Model 1(Basic): Constant reproduction number and poisson offspring distribution

2. Model 2(NegBin): Constant reproduction number and negative binomial offspring distribution

3. Model 3(Varrying Reproduction Number): Varrying eproduction number and poissonoffspring distribution

4. Model 4(Full): Varrying reproduction number and negative binomial offspring distribution

We will assess how well the models work both for the entire epidemic and for each health zone. To evalute a model we will estimate the calibration, sharpness, bias, log score and CRPS for forecasts of 1, 7, 14, 21 and 28 days ahead. To do this we start **check** 16 days after the start of the epidemic in the location and calculate the $d$ ahead prediction for all historically available data. For the calibration we use all the values to assess if they are normally distributed, while for all the other metrics we average them over all the time steps. 16 days was chosen as the start as this is when the method for calculatng $R_t$ can give somehat reliable values [12].

# 5 Results

From the start of the outbreak until XX.XX.XXXX there has been XXXX confirmed ebola cases and XXXX deaths. Figure 1 shows the weekly number of cases and Figure 2 the total number of cases from each Health Zone. From Figure 1 and the estimte of the instantaneous reproduction number in Figure 3 we can see the evolution of the outbreak. After the initial period were there were no daily surveillance, there was an increase in cases from October 2018 with a large reproduction number, followed a more varied period where the reproduction number varied around 1. From March 2019 the number of cases per week increased signficantly, with a continuing large number of cases in June and July even if the reproduction number seems to decrease.
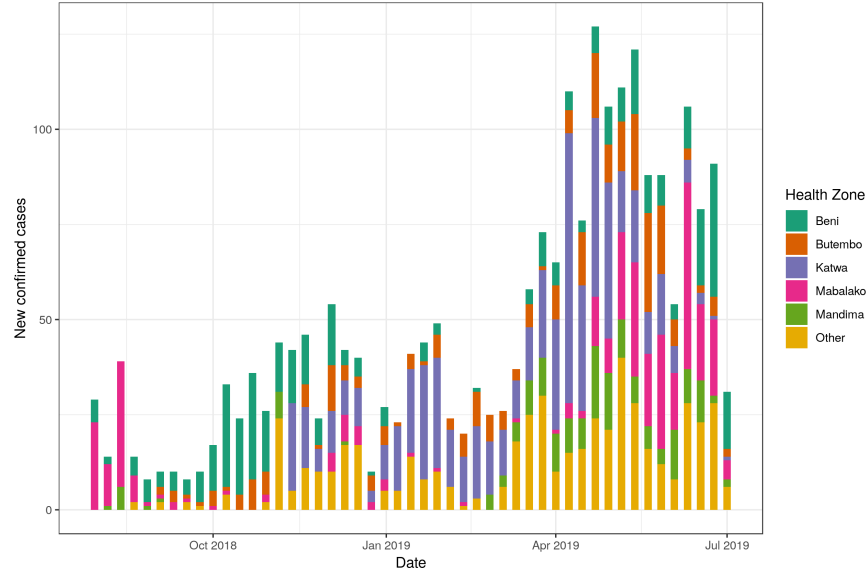
# 6 Discussion

# 7 Conclusions

Figure 1: Number of new confirmed cases by week in the health zones with the most Ebola Cases

# References

[1] J.-P. Chretien, S. Riley, and D. B. George, "Mathematical modeling of the West Africa Ebola epidemic," *eLife*, vol. 4.

[2] A. Camacho, A. Kucharski, Y. Aki-Sawyerr, M. A. White, S. Flasche, M. Baguelin, T. Pollington, J. R. Carney, R. Glover, E. Smout, A. Tiffany, W. J. Edmunds, and S. Funk, "Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-time Modelling Study," *PLoS Currents*, vol. 7, Feb. 2015.

[3] C. Rivers, J.-P. Chretien, S. Riley, J. A. Pavlin, A. Woodward, D. Brett-Major, I. M. Berry, L. Morton, R. G. Jarman, M. Biggerstaff, M. A. Johansson, N. G. Reich, D. Meyer, M. R. Snyder, and S. Pollett, "Using outbreak science to strengthen the use of models during epidemics," *Nature Communications*, vol. 10, pp. 1–3, July 2019.

[4] W. H. Organization, "Ebola Outbreak in DRC 01: 7 August 2018," Aug. 2018.

[5] W. H. Organization, "Ebola Outbreak in DRC 09: 04 October 2018," Oct. 2018.

[6] W. H. Organization, "Ebola Outbreak in DRC 16: 21 November 2018," Nov. 2018.

[7] W. H. Organization, "Ebola Outbreak in DRC 38: 24 April 2019," Apr. 2019.

[8] W. H. Organization, "Preliminary results on the efficacy of rVSV-ZEBOV-GP Ebola vaccine using the ring vaccination strategy in the control of an Ebola outbreak in the Democratic Republic of the Congo: an example of integration of research into epidemic response.," tech. rep., Apr. 2019.
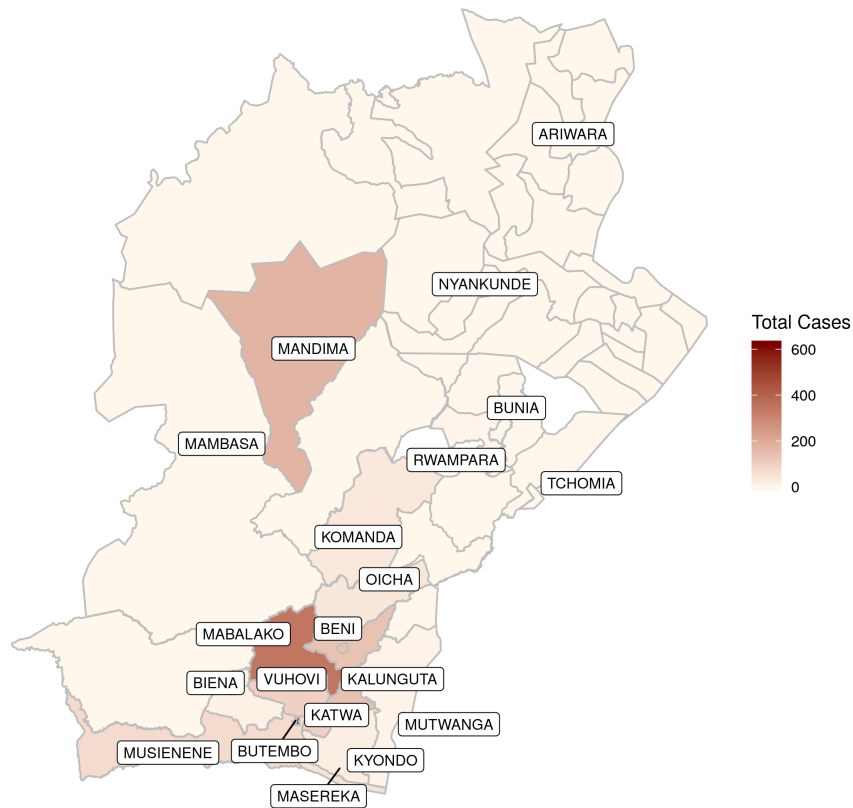
Figure 2: Total number of confirmed cases by Health Zone

[9] N. I. of Allergy and Infectious Diseases, "Independent Monitoring Board Recommends Early Termination of Ebola Therapeutics Trial in DRC Because of Favorable Results with Two of Four Candidates," Aug. 2019.

[10] W. H. Organization, "Ebola Outbreak in DRC 51: 23 July 2019," July 2019.

[11] C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, and A. Vespignani, "The RAPIDD Ebola Forecasting Challenge: Synthesis and Lessons Learnt," *Epidemics*, vol. 22, pp. 13–21, Mar. 2018.

[12] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez, "A new framework and software to estimate time-varying reproduction numbers during epidemics," *American Journal of Epidemiology*, vol. 178, pp. 1505–1512, Nov. 2013.

[13] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, "Superspreading and the effect of individual variation on disease emergence," *Nature*, vol. 438, p. 355, Nov. 2005.

[14] P. Nouvellet, A. Cori, T. Garske, I. M. Blake, I. Dorigatti, W. Hinsley, T. Jombart, H. L. Mills, G. Nedjati-Gilani, M. D. Van Kerkhove, C. Fraser, C. A. Donnelly, N. M.
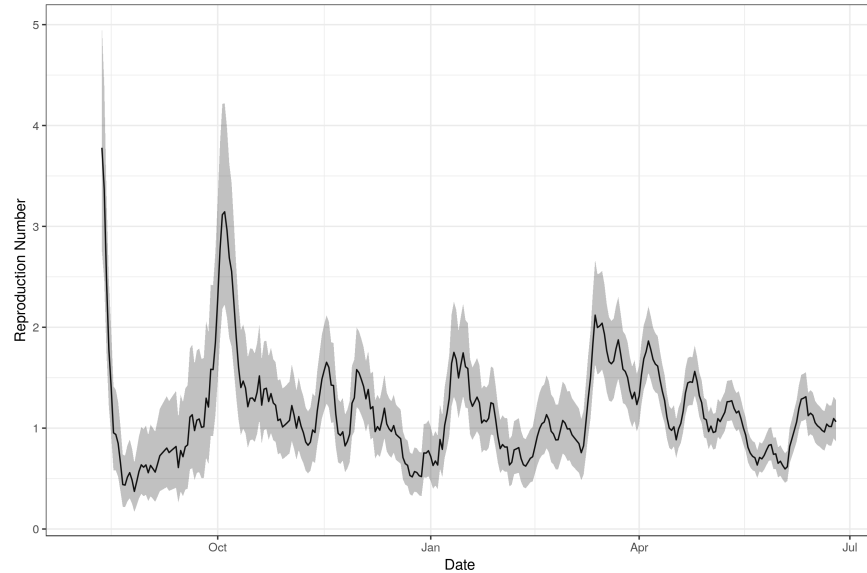
Figure 3: Reprodcution number as a function of time, the shaded region shows the 95% credible interval

Ferguson, and S. Riley, "A simple approach to measure transmissibility and forecast incidence," *Epidemics*, vol. 22, pp. 29–35, Mar. 2018.

[15] "Ebola Virus Disease in West Africa  The First 9 Months of the Epidemic and Forward Projections," *New England Journal of Medicine*, vol. 371, pp. 1481–1495, Oct. 2014.

[16] International Ebola Response Team, J. Agua-Agum, A. Ariyarajah, B. Aylward, L. Bawo, P. Bilivogui, I. M. Blake, R. J. Brennan, A. Cawthorne, E. Cleary, P. Clement, R. Conteh, A. Cori, F. Dafae, B. Dahl, J.-M. Dangou, B. Diallo, C. A. Donnelly, I. Dorigatti, C. Dye, T. Eckmanns, M. Fallah, N. M. Ferguson, L. Fiebig, C. Fraser, T. Garske, L. Gonzalez, E. Hamblion, N. Hamid, S. Hersey, W. Hinsley, A. Jambei, T. Jombart, D. Kargbo, S. Keita, M. Kinzer, F. K. George, B. Godefroy, G. Gutierrez, N. Kannangarage, H. L. Mills, T. Moller, S. Meijers, Y. Mohamed, O. Morgan, G. Nedjati-Gilani, E. Newton, P. Nouvellet, T. Nyenswah, W. Perea, D. Perkins, S. Riley, G. Rodier, M. Rondy, M. Sagrado, C. Savulescu, I. J. Schafer, D. Schumacher, T. Seyler, A. Shah, M. D. Van Kerkhove, C. S. Wesseh, and Z. Yoti, "Exposure Patterns Driving Ebola Transmission in West Africa: A Retrospective Observational Study," *PLoS medicine*, vol. 13, p. e1002170, Nov. 2016.

[17] A. Cori, *EpiEstim: EpiEstim: a package to estimate time varying reproduction numbers from epidemic curves*. 2013.

[18] S. L. Scott, *bsts: Bayesian Structural Time Series*. 2019.

[19] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 243–268, 2007.

[20] S. Funk, A. Camacho, A. J. Kucharski, R. Lowe, R. M. Eggo, and W. J. Edmunds, "Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15," *PLOS Computational Biology*, vol. 15, p. e1006785, Feb. 2019.

[21] C. Czado, T. Gneiting, and L. Held, "Predictive Model Assessment for Count Data," *Biometrics*, vol. 65, no. 4, pp. 1254–1261, 2009.

[22] J. Faraway, G. Marsaglia, J. Marsaglia, and A. Baddeley, *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*. 2017.

[23] A. Jordan, F. Krueger, and S. Lerch, "Evaluating Probabilistic Forecasts with scoringRules," *Journal of Statistical Software*, 2018.

[24] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.