TITLE

CANDIDATE NUMBER

August 4, 2019

Word count: XXXX

Abstract

Abstract

Contents

1 Introduction **2**

2 Background **2**

3 Methods **2**

3.1 Model 2

3.1.1 Offspring distribution 3

3.1.2 Reproduction number 3

3.1.3 Simulating from the model 4

3.2 Assessing probabilistic forecasts 5

3.3 Implementation 6

4 Results **6**

5 Conclusions **6**

1 Introduction

2 Background

3 Methods

3.1 Model

We consider a model where the daily incidence, I_t follows a branching process. Similar models have been considered in [1, 2] and have been used to modell Ebola during the 2014-2016 Ebola outbreak in Wester Africa [3, 4]. In this class of models each infectious person gives rise to ν new infections, where ν 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 force of infection λ_t. The force of infection is a producted of previous incidence weighted by the serial interval w_τ 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 [3]. 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.

3.1.1 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 λ. 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 [2]. 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, μ 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 $k = 0.16$[2]. For the Ebola outbreak in West Africa they found $k = 0.53$ using conservative assumptions and data from the network of infections [4].

PLOT of Distributions?

The distribution for I_t would be give 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.

3.1.2 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. The model could for example reproduce in expetation a standard compartmental model where the reproduction number is given by $R(t) = R_o s(t)$, where $s(t)$ is the fraction of susceptibles. In a normal SIER model with a constant rate to move between the "E" and the "I" compartment the serial interval woudl be exponential distributed.

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 [1]. 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 [5] to estimate the reproduction number.

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 [1] 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 $\log(R_t)$. We use model with a semi-local linear trend to allow the estimation of a trend in the recent data. We use $\log(R_t)$ instead of R_t to ensure that the reproduction number remains positive. We use the R-package BSTS [6] and the standard priors. The model we use is specified as follows

$$log(R_{t+1}) = log(R_t) + \delta_t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\mu),$$
$$\delta_{t+1} = D + \phi(\delta_t - D) + \eta_t, \eta_t \sim N(0, \sigma_\delta).$$

δ_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 σ_μ and σ_δ, a gaussian prior on D and a gaussian prior for ϕ. 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 $log(R_t)$ with uncertainties.

3.1.3 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 λ_{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 λ_{t+2} we use the sampled value for I_{t+1} together with the historical data I_s for $s \leq t$.

3.2 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 [7]. In addition we will consider proper scoring rules for comparing probability ditributions. We follow the approach taken in [8], 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) [9]. 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 ν 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 [10]) 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.

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 logarithmic and the contineously ranked probabilty score(CRPS) as implemented in the scoringRules package [11]. The logarithmic scoreing rule is given by

$$logS(F_t, k_t) = -log(F_t(k_t),$$

and the CRPS score as

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

To evalute both of these scores from simulated samples a kernel density estimate is used to estimate F_t from the samples.

3.3 Implementation

The model was implemented in the R programming langauge [12] 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 [1].

4 Results

5 Conclusions

References

[1] 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.

[2] 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.

[3] “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.

[4] 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.

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

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

[7] 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.

[8] 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.

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

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

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

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