Forecasting Infectious Disease Epidemics

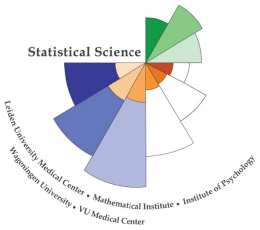
Laura Verkerk (s0906646)

Thesis advisor: Prof. Dr. J. Wallinga

Second thesis advisor: Prof. Dr. H. Putter

MASTER THESIS

Defended on June 29, 2018



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Contents

Abstract

In this project a new approach to forecasting infectious disease epidemics was tested in a simulation and applied to data of the 2014 - 2016 Ebola epidemic. GLMs were applied to the (simulated) data, from which the key quantities contact rate and epidemic size could be obtained. With (non-)parametric bootstrapping, the GLM results could be assessed, and the key quantities were obtained and subsequently used to produce forecasts. Forecasting intervals were made to show the accuracy of the forecasts in terms of epidemic size and duration. Simulation results suggested that the method underestimated the eventual epidemic size, and overestimated the contact rate. However, applying the method to a real-life data set resulted in overestimation of the eventual epidemic size. The results of the contact rate for the application on real-life data should be compared to estimates from literature, before a significant meaning can be given to the results. Both simulation and application results gave variable estimates for the epidemic duration, although a positive relation was seen between epidemic size and epidemic length. Estimates for the contact rate could be improved. The major issues with prediction were accountable to exact collinearity introducted by the systematic model; the major issues with forecasting were accountable to extreme estimates of the epidemic size. The cause of both issues lies in the GLMs that were fit to the data.

Word of thanks

First and foremost, I would like to thank my thesis supervisors, Jacco Wallinga and Hein Putter, for their unending positivity, patience, and the many hours they have spent guiding me in this project. The past few years have been very difficult due to personal circumstances, and their attitude in guiding me has been a beacon of rest and motivation in the last year. I have constantly admired the enthusiasm Jacco and Hein have for their work, research and this project, and their positive attitude in life. Gentlemen, I wish you the best, stay as you are.

Secondly, I would like to thank the whole body of Statistical Science, which gave an excellent master's education, which learned me to work really hard, and which provided me with considerable skill to confidently start in the job market. The coordinating committee showed understanding and patience for my situation, and I am very grateful that I have been given the chance to continue my education on a pace with which I could cope.

Thirdly, I would like to thank my friends and family for their unwavering support, attention, motivational talks, understanding and love. I am very lucky to have so many great people around me.

1. Introduction

Accurate forecasts are essential to the control of infectious disease epidemics, but epidemiologists have not yet succeeded in finding a proper method. In particular, early forecasts of important information such as peaks in incidence and spatial spread of the infection would enable decision makers to act sooner and minimize the damage to society once a disease outbreak is detected.

The forecasting of infectious disease epidemics has been under widespread attention worldwide, and continues to pose a challenge to researchers. The 2014 ebola disease epidemic in Africa (Aywlard et al, 2014) has lead to several attempts to model the infection severity and spread accurately, even in the form of contests issued by governmental instances. An example of a recent competition is the 'RAPIDD Ebola Forecasting Challenge', hosted by the Research and Policy for Infectious Disease Dynamics (RAPIDD) group of the American National Institutes of Health (NIH) (http://www.ebola-challenge.org/). An analysis of the 8 models in this competition was performed by Viboud et al. (2017). They showed that for short-term forecasting, locally fitted adaptive models performed best. Overall, predictions based on a Bayesian average of the results of the participating models in this study, outperformed any individual one, indicating that there might not be one sure way to victory in the case of forecasts.

Other infectious diseases, for example influenza, have been an ongoing topic of interest as well. The annual epidemics of influenza caused the American Centers for Disease Control and Prevention (CDC) to hold a challenge to predict Influenza epidemics. The goal is to make the most accurate forecast for that year's influenza wave. The teams can use any available data source.

Ong et al. (2010) showed that real-time epidemic activity can be monitored relatively easy, by means of reporting of influenza-like illness (ILI) from general practise or family doctor clinics on a daily basis. Their real-time surveillance method showed that the progress, peak and end of an epidemic can be forecasted by daily refitting a stochastic model of disease dynamics using particle filtering.

Infectious disease epidemics are often modeled with dynamical models. Dynamical models deal with processes which vary over time. A simple example would be exponential growth: one group of the Ebola Challenge used an exponential growth model (Aylward et al., 2014). An overview of Ebola models and the available data is presented in a recent article by Backer and Wallinga (2016). Here we are dealing with time series of incidence of infections (Gandon et al., 2016; Wood, 2010; King et al., 2008).

An alternative to data-based forecasting has recentely been proposed by David Farrow and collaborators (2017). They propose the use of the web-based forecasting system 'Epicast', in which human judgement is compared to data-driven forecasts. They hypothesized that humans are able to assimilate information from several sources relatively easily, and that the 'Wisdom of the crowd' often creates a robust answer given that the underlying distribution is unbiased. However, using Epicast consistently during the influenza season posed a challenge to many participants. Additionally, human forecasts need to be done in real time, while data-driven approaches can be applied retrospectively.

Aim

In this project we investigated a statistical approach towards predicting and forecasting epidemics. As prediction is a familiar field in statistics, we hoped to use this to our advantage. Although the method is familiar to most statisticians, it has not yet been used to predict the development of infectious diseases.

Method

Time-to-event data was modeled using a generalized linear model (GLM) with the infection events as outcome variable. The approach was first tested using simulated data sets, to which the model was applied in several situations, such as binned data and underreporting. Incubation time was automatically included in the generation of simulated data sets. Forecasts were created from halfway through the epidemic, as forecasts are usually made during the development of an epidemic. The aim was to retrieve the infection rate and the

final number of infection events in an epidemic. Secondly, the approach was applied to data on the Ebola epidemic in West Africa, 2014.

Expected results

Several statistics were obtained from applying the method. First of all, predictions and prediction intervals were acquired so that model quality could be assessed. Secondly, forecasts and forecasting intervals were created, showing the predictive value of the models. We expect that this method will give somewhat accurate forecasts, as the natural form of an epidemic was used. However, models made at the first few days of an epidemic will probably give broader forecasting intervals and worse fitting models than models made halfway through the epidemic. It is expected that the original infection rate and final number of infection events in an epidemic can accurately be obtained by means of bootstrapping.

Future implications

More accurate prediction and forecasting of infections will allow decision makers and health care instances to respond more adequately to upcoming epidemics.

Set-up of the thesis

The thesis is divided into the stages of the simulation, describing for each stage first the method, then the results, finishing with a last short discussion for each section. Section 2 describes the data generation process for the simulated data used in this study. Section 3 shows the generalized linear models that were used, and a simple example is given in which the models were applied to one simulated data set. Section 4 describes the process of obtaining the infection rate and epidemic size. In Section 5 bootstrapped predictions were made for simulated time-to-event data sets. Section 6 shows bootstrapped forecasts for the same data sets. Section 7 shows a case study on ebola. Section 8 gives an overall discussion of the results and recommends future research.

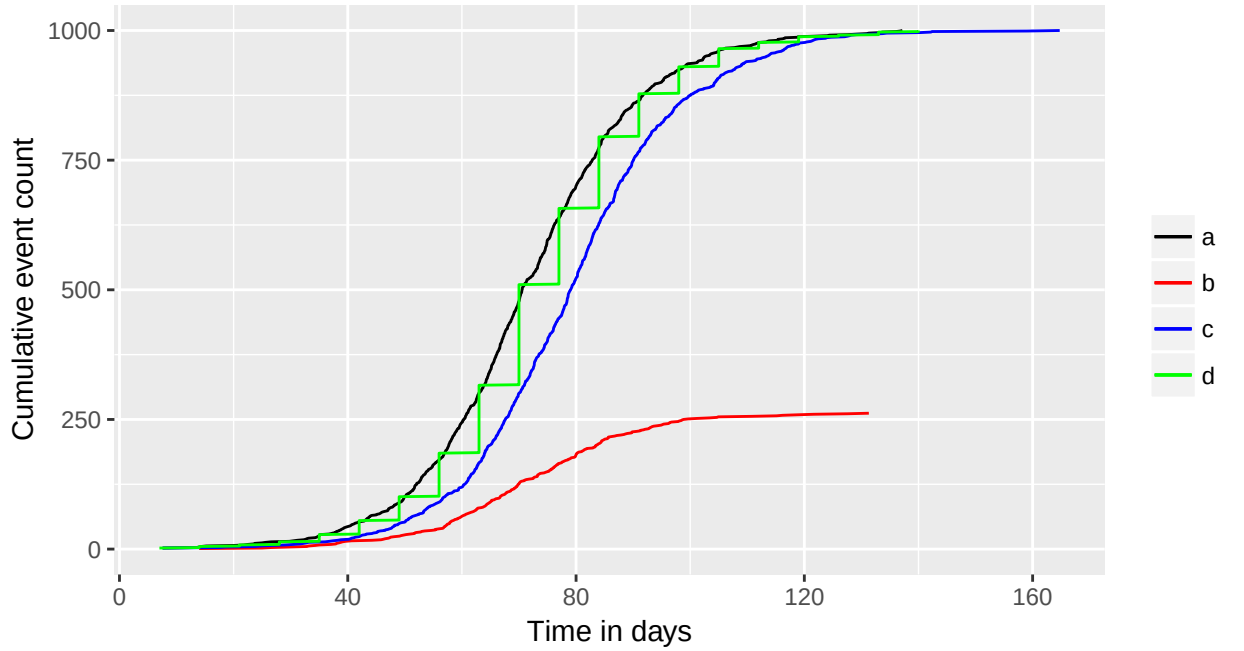


Figure 2.1. Cumulative Infection plot for time in days (x-axis) versus cumulative number of infections (y-axis)

2. Data simulation

Data sets were simulated to resemble an infectious disease epidemic. The goal was to predict the number of infections in the future, which was be done by seeing the epidemic as a time-to-event data set where the outcome is the time between successive events. The *symptom onset* is the moment an individual starts experiencing symptoms of the infection. The *incubation period* is the time between infection and symptom onset. These key quantities are essential for a proper prediction and were included in the simulation.

The basic algorithm, which is described in Section 2.1, simulates an epidemic with bounds on prevalence of infection: all people who are not infected yet, are still susceptible to infection, and the algorithm continues until all people in the population are infected, or to a predefined maximum such as 10% of the infection. The first extension of this algorithm, as described in Section 2.2, adds underreporting to the simulation. The second extension of this algorithm can be seen in Section 2.3, where incubation time is introduced: the time between infection and symptom onset. As infectious diseases are usually recorded on a daily, weekly or monthly basis, Section 2.4 includes binning of the events into time windows.

To illustrate the simulations, figures were produced for a single simulation run. **Figure 2.1** shows simulated infections for each case described, with the time in days on the x-axis and cumulative number of infections (event count c) on the y-axis: the black line (a) reflects the basic algorithm, the red line (b) shows underreporting, the blue line (c) shows incubation time, and the green line (d) shows binning. **Figure 2.2** shows the time-to-event data that can be created from the infection data. Respectively, panels **a, b**, and **c** show the basic algorithm, underreporting and incubation time as described above. To make time-to-event data from a binned data set, the data needs to be jittered first. As such, **d** shows binned and subsequently jittered data. On the x-axis is the cumulative event count and on the y-axis the time (in days) is shown.

2.1 An epidemic with bounds on prevalence of infection

An overly simplified infectious disease epidemic was simulated. The epidemic starts with one infected individual, corresponding to a state $I = 1$. Then a contact rate β is defined, which describes the number of

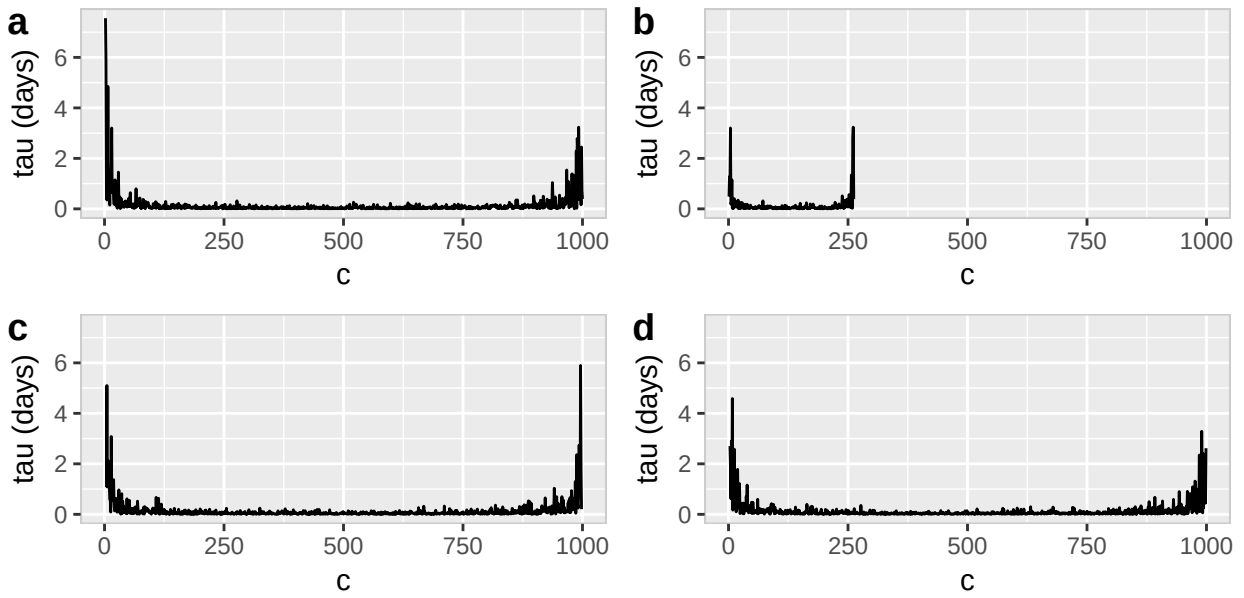


Figure 2.2. Time-to-event plots between subsequent infections. The number of infected individuals can be seen on the x-axis and the time until the next infection on the y-axis. Panel (a) shows an epidemic made with the basic algorithm, (b) an epidemic with underreporting, (c) an epidemic with incbation time, and (d) a binned and jittered epidemic.

infectious contacts an infected individual has during the course of a day. Days were chosen here, and binning this into other time units such as weeks or months could be added later. The size N of the population was predefined and we assume that all uninfected subjects are susceptible at all times with a specific probability, $1 - \frac{I-1}{N-1}$. The effective infection rate λ is defined as the product of a) the rate of contacts per day , b) the number of infectious individuals, and c) the susceptibility probability: $\beta I(1 - \frac{I-1}{N-1})$. The state at time t is $I(t)$, and the effective infection rate at time t is $\lambda(t)$. The events -a new infection- are assumed to occur according to a Poisson process, where the time to the next event, τ, is given by an exponential distribution with probability density function

$$P(\tau|\lambda(t)) = \lambda(t)e^{-\lambda(t)\tau}.$$

Time to the next event is always positive: as such the exponential distribution, which is a variant of the gamma distribution, was appropriate to use here. The distributional properties of the gamma distribution allow for extra flexibility in modeling the time-to-event data. The infection curve corresponding to this epidemic can be seen in **Figure 2.1(a)**. The infection grows slowly at the beginning and at the end of the epidemic: the effective infection rate λ is very low when few people are infected. The infection then grows exponentially until half of the population is infected and the other half susceptible, to enter a declining growth when this turning point is passed. **Figure 2.2(a)** depicts the time-to-event data. The time-to-event is large at the beginning and end of the epidemic and small towards the middle, creating a bathtub shaped plot. This plot corresponds with the effective infection rate.

2.2 Underreporting

In 2.1 every infection was immediately detected: in real life, not all infections are reported. Even worse, with some viruses such as influenza, the reporting rate, p, is low: in the Netherlands only 20% of people *infected* with influenza-like symptoms consults a doctor (Friesema et al., 2009). To simulate underreporting, a Bernoulli trial is done for each event in the data. For each simulated infection event a random number is drawn

from a uniform(0,1)-distribution. When the number is below the assumed probability p that an individual ends up in the surveillance system, the observation will be omitted. **Figure 2.1(b)** shows underreported data with a reporting rate of $p = 0.25$. Surprisingly, even with only 25% of the data available, the curve of the original infection is followed closely. The underreported data creates a less strong bathtub shape, as can be seen in **Figure 2.2(b)**. It can also be seen that the largest time-to-event is about 3-4 days in the tails of the underreported epidemic, whereas the fully reported epidemics can peak to 6 days.

2.3 Symptom onset

There is some time between infection and the onset of symptoms: the incubation time. The symptom onset is when people enter the surveillance system, and usually the symptom onset is recorded instead of infection time, given some exceptions as in the study of (Aylward et al., 2014), where they interrogated all patients about the contacts they have had in the past few weeks, to trace back the actual infection time. For the data simulation it is assumed that individual incubation times are independent and distributed according to a gamma distribution, with a mean of 10 days and a standard deviation of 2 days. This corresponds to a shape α of 25, and a rate β of 2.5. The time until the next event, τ, is then given by adding the incubation time t_{incub} to the time point of infection t. A constraint in this set-up is the assumption that a new infection can only take place when the symptoms are detectable. **Figure 2.1(c)** shows that the infection curve is shifted to later in time and that it is slightly stretched. In **Figure 2.2(c)** it can be seen that the tails of the bathtub are thicker, indicating a longer time between sequential symptom onsets. This shows that the incubation time is being added to each infection time, as discussed previously in this paragraph.

2.4 Binning

Recording of individual infections is often not done: the number of infections is recorded on a daily, weekly, monthly or even yearly basis. To simulate such data, binning was included in the algorithm by replacing the time variable, t, by an index of the time bin in which the corresponding observation should be. By grouping the simulated data into counts per bin, several types of generalized linear models could be applied, which will be described in Section 3. In the simulation, the specification of days, weeks, months, years, decades and any other number of days was allowed. The binned infection curve in **Figure 2.1(d)** shows a blocked structure with jumps between bins. A time-to-event plot for binned data was not made, as it would result in many zeroes, with an occasional jump of one time-unit when the bin shifts.

However, **Figure 2.2(d)** does show the result of a binned data set: suppose we would want to analyze binned data, there are several options available. The data could be jittered, stretching the data from weeks to individual observations with their own respective event times: random numbers drawn from a uniform(0,1)-distribution are added to the bin time, and subsequently the jittered time is transformed to days. Every event gets an unique time point, allowing for the creation of a time-to-event data set, which can be seen in **Figure 2.2(d)**. The jittered data can be analyzed the same way as unbinned data sets.

Another option is to leave the data as is: event counts in every bin could be assumed to be created according to a Poisson process, and the data could be analyzed with a method fit for counts. However, when we want to stay in terms of time-to-event data, counts are not sufficient.

Another way of dealing with binned data would be by assuming that, within each bin, the observations are equally spread across the time in this bin. In a way, it is similar to applying weights to the bins with weight size corresponding to the number of events within this bin. Though this seems similar to the Poisson process proposed earlier, the way the data is regarded is *not* as a process where the counts per bin are the observations, but rather, as a process where the bins are the observations, weighted by the density of observations in this bin. We will discuss the processes of jittering and weighting of binned data in section 3 on GLM.

2.5 Variations in the simulation basics

There are many things that can be varied in a simulation research. A summation is given of the possibilities and the variations that were used in our research are mentioned here. The following variations could be used for the simulated data sets:

- β: the infection rate. A large β creates a quickly advancing epidemic. A higher infection rate β will accelerate the epidemic, but will not necessarily change its shape;
- N: the population size: a larger population size will give more reliable results;
- μ: the mean of the gamma distribution for incubation time;
- σ: the standard deviation of the gamma distribution for incubation time;
- $Imax$: setting a maximum number of infections can be useful when a specific threshold is required to start e.g. mass cure programs;
- $bins$: which bin is used for this data: day, week, month, year, decade, or a random number of days;
- p: the reporting rate could be used to draw conclusions about the possible loss of information as a result of underreporting during an epidemic.

In this study, the incubation time was defined to be a gamma distribution with mean $\mu = 10$ and standard deviation σ of 2 days. For now, the population size N was set at $N = 10000$ for the example in section 3, and at $N = 1000$ for the bootstrapped simulations in section 5 and 6. It is assumed that the progression of an infection will have the same shape regardless of the population size. Initially, the infection rate β was varied to see the progress of the epidemic. It was fixed at $\beta = 0.1$ for further analysis, as a higher infection rate β did accelerate the epidemic, but did not necessarily change its shape. The reporting rate p was set at $p = 1$ and at $p = 0.25$ for section 3.

3. Regression

To analyze the time-to-event data, several generalized linear models (GLMs) were fit to the data. The gamma family was chosen, as the time-to-event data in this study was assumed to have an exponential distribution: the exponential distribution is a special case of the gamma distribution with shape parameter $\alpha = 1$. Additionally, a GLM was fit with the Poisson family as the Poisson is appropriate for analyzing counts in e.g. binned data.

We start with explaining the generalized linear model. A vector $\boldsymbol{y}$ of length n is observed of which the elements are assumed to be realizations of the random variable $\boldsymbol{Y}$ with means $\boldsymbol{\mu}$. A generalized linear model consists of three parts (McCullagh & Nelder, 1989).

First, the *random* component is the random variable $\boldsymbol{Y}$, of which the elements are identically, independently distributed according to some distribution with means $\boldsymbol{\mu}$. In classical linear models the distribution of $\boldsymbol{Y}$ is assumed to be Gaussian; however, in generalized linear models the distribution of $\boldsymbol{Y}$ is allowed to be from any exponential family, including the Gaussian.

Second, the *systematic* component describes the vector $\boldsymbol{\mu}$ in terms of some unknown parameters $\beta_1, ..., \beta_p$ that can be estimated from the data. The covariates $\boldsymbol{x}_1, ..., \boldsymbol{x}_p$ produce a linear predictor $\boldsymbol{\eta}$ given by

$$\boldsymbol{\eta} = \sum_{1}^{p} \boldsymbol{x}_j \beta_j.$$

In all applied GLMs, the systematic component $\beta_0 + \beta_1 x + \beta_2 x^2$ was used to model the data, where x is the cumulative number of events and x^2 is the squared cumulative number of events.The β's are the regression coefficients.

Third, the *link* connects the random and systematic components: it relates the linear predictor η to the expected value μ of a data point y with a *link function* $g(\mu) = \eta$. The classical linear model can be seen as a GLM with Gaussian data and an identity link $\eta = \mu$. A link function can be extended to any monotonic differentiable function. For this thesis the inverse link $\eta = 1/\mu$ will be used for the gamma GLM and the identity link will be used for the Poisson GLM. The inverse link is used in the gamma GLM because the time-to-event data has a bathtub shape. The inverse link connects this bathtub shape, as seen with the data simulation, to the original exponential distribution of the epidemic. For the Poisson GLM the identity link was used to keep comparability to the gamma GLM.

The regression coefficients β_0, β_1 and β_2 are estimated by the method of maximum likelihood. For generalized linear models, the maximum likelihood is expressed as follows: suppose we have a density function $f(y; \theta)$ for an observation y given the parameter θ, then the log likelihood, expressed as a function of the mean-value parameter, $\mu = E(Y)$, is

$$l(\mu; y) = \log\, f(y; \theta).$$

For a set of independent observations $\boldsymbol{y} = y_1, ..., y_n$ the log likelihood is the sum of the individual contributions

$$l(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i} \log\, f_i(y_i; \theta_i)$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ is a vector of means.

Before describing in which fashion the log likelihood might be used for fit statistics, and before discussing issues related to these fit statistics such as the dispersion parameter, it is important to know which GLMs were applied in this study, including some details on these particular GLMs. The three GLMs applied were the gamma GLM, weighted gamma GLM, and Poisson GLM, which will be discussed first.

gamma GLM

The infection data was simulated according to an exponential distribution, with an added gamma distribution in the case of incubation time. The gamma GLM with an inverse link was applied to the data. The random

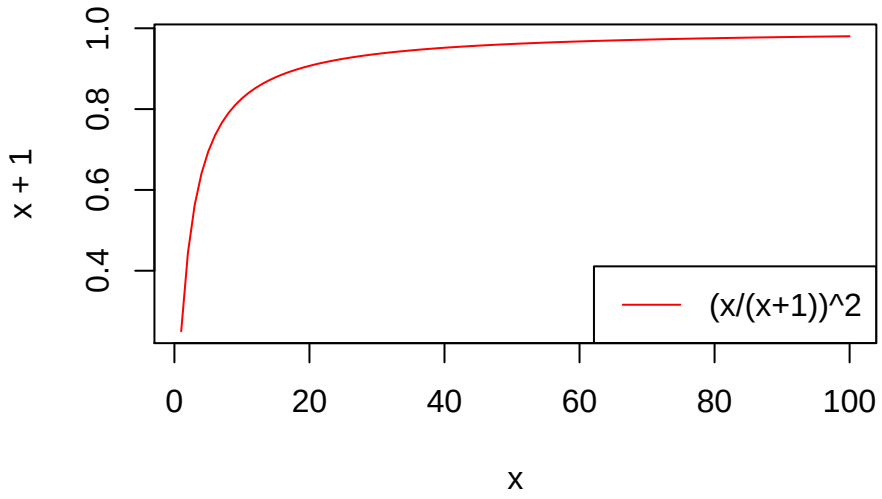


Figure 3.1. Relative change in y for the Gamma GLM between two subsequent values of x. The red line represents y.

component, $\boldsymbol{Y}$, has a $\Gamma(\alpha, \beta)$-distribution with probability density function:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

The systematic component is taken to be $\eta = \beta_0 + \beta_1 x + \beta_2 x^2$, where x is the cumulative number of events and x^2 the quadratic cumulative number of events. The inverse link function $\eta = 1/\mu$ was used to connect the systematic and random component:

$$\mu = \frac{1}{\beta_0 + \beta_1 x + \beta_2 x^2}.$$

When looking at regression coefficients for a classical linear model, a one-unit change in x would change the outcome y with β_1, given that the other variables are kept constant. One would generally assume that the covariates are unrelated. However, in this study the covariates x and x^2 are related. Suppose that $\mu_x = \frac{1}{\beta_0+\beta_1 x+\beta_2 x^2}$, then the next value, μ_{x+1}, would amount to $\mu_{x+1} = \frac{1}{\beta_0+\beta_1(x+1)+\beta_2(x+1)^2}$. When β_0 and β_1 are not too big, and β_2 not too small, the relative change in y would approximately be

$$\frac{\mu_{x+1}}{\mu_x} = \frac{\beta_0 + \beta_1 x + \beta_2 x^2}{\beta_0 + \beta_1(x+1) + \beta_2(x+1)^2} \approx (\frac{x}{x+1})^2.$$

When plotting this, an exponential pattern shows, as can be seen in **Figure 3.1**

The gamma GLM was used for both binned and unbinned data. For binned data, jittering was necessary before applying a gamma GLM.

weighted gamma GLM

One of the distributional properties of the gamma GLM is summation: provided that the distributions are independent, they can be summed when the shape parameter α is equal across these distributions. The result

of summing n observations in a bin with duration y (or Δt) and mean μ is a gamma distribution with shape parameter $n\alpha$ and mean $\mu_1 + ... + \mu_n$.

The weighted gamma GLM was used slightly different than the gamma GLM. First of all, the weighted gamma GLM was applied to data at bin level. Second, the response variable y was transformed to, and used as the average time to event per bin, y^w.

Suppose that y_i is the total time to event from n_i observations in bin i:

$$y_i = \sum_{j:1}^{n_i} y_{ij},$$

where y_{ij} is the j^{th} time to event in bin i. Then,

$$y_i^w = y_i/n_i$$

is the average time to event in bin i. If $y_{ij} \sim gamma(\mu_i, \alpha)$ is independent, then

$$y_i^w \sim gamma(\mu_i, n_i\alpha),$$

where α is the shape parameter.

This corresponds to assigning a weight of n_i to an observed mean duration y_i^w. This weight can be implemented in the GLM algorithm by assigning a vector of weights to the binned observations, and by dividing the binned time variable by the number of observations in that bin.

Poisson GLM

The Poisson GLM can be used to analyze counts or rates. In the case of infectious disease epidemics, the data is often binned on a week level with the number of new infections as recorded information. Poisson GLM is easiest to apply from the three approaches for binned data in this study: the data is analyzed at bin level, without any additional actions. The predictor x is the cumulative amount of infected individuals, x^2 the square of x, and the response variable y is the number of infected individuals in each bin.

A Poisson GLM with an identity link was applied to the data. The random component, $\boldsymbol{Y}$, is described for parameter λ with the Poisson distribution: $Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$. Identical to the gamma GLM, the systematic component is $\beta_0 + \beta_1 x + \beta_2 x^2$, where x is the cumulative number of events and x^2 is the quadratic cumulative number of events. The identity link function $\eta = \lambda$ was used to connect the systematic and random components:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2.$$

In this section, four GLMs were applied *once* to the simulated data to illustrate the methods: twice the gamma GLM, once on originally simulated time-to-event data and once to jittered data; once the weighted gamma GLM on binned data, and once the Poisson GLM on binned data. Additionally, to illustrate underreporting, these four GLMs were applied to data with a reporting rate of 25%, creating eight models in total. In Section 5 bootstrapped estimates are presented for these eight models. Here we continue with discussing the deviance and dispersion, to subsequently discuss the results of applying the eight models to simulated data.

Deviance

In the context of GLMs the *deviance* is often used as the goodness-of-fit criterion in stead of the log likelihood. The deviance function measures the discrepancy between the data vector $\boldsymbol{y}$ and the mean vector $\boldsymbol{\mu}$. McCullagh and Nelder (1989) defined the deviance as a linear function:

$$D(\boldsymbol{y}; \boldsymbol{\mu}) = 2\phi\{l(\boldsymbol{y}; \boldsymbol{y}) - l(\boldsymbol{\mu}; \boldsymbol{y})\},$$

where $l(\boldsymbol{y};\boldsymbol{y})$ is the maximum likelihood that can be obtained for an exact fit when the observed data is equal to the fitted values, and ϕ is the *dispersion parameter*.

The dispersion parameter is 1 if the variation in the data is exactly as large as would be expected based on the proposed statistical model. *Overdispersion* indicates that there is more variation in the data than would be expected based on the proposed statistical model. McCullagh (1992) stated that the deviance estimate does not depend on ϕ, as the estimation of the regression coefficients $\boldsymbol{\beta}$ is independent of ϕ: in all generalized linear models, the dispersion parameter ϕ is regarded as either known or else an unknown constant independent of the covariates. The dispersion is either fixed at 1 or can be estimated from the residuals: McCullagh and Nelder (1989, p.p. 30) show an overview. When the Poisson family is specified, the dispersion parameter is fixed at one. For the Gamma GLM, the dispersion is estimated as α^{-1}, where α is the shape parameter. Checking for overdispersion can give valuable insights in the performance of a model: in the case of overdispersion, it might be useful to specify a more complicated model.

Nelder and Wedderburn (1972) defined the deviance as minus twice the maximized log-likelihood for a given model, $-2L_{max}$. The maximized log-likelihood for a given model is the optimal trade-off between model complexity and the matching of the model to the data.

The least complex model is the null model, which assumes that there is one parameter, representing a common μ for all ys, to estimate for all data points. In the null model, all variation between the ys is due to the random component of the GLM. The most complex model is the saturated model. In this model, each data point has its own parameter. For the saturated model as many parameters are estimated as there are observations: all the variation in the ys is assumed to come from the systematic component of the GLM, leaving none for the random component.

Two types of deviance were used to assess model fit: the null deviance and the residual deviance. The null deviance is the deviance between the saturated model and the null model. The residual deviance is the deviance between the model we want to fit and the saturated model. A small residual deviance indicates that the model fits the data quite well.

For the regular, jittered and weighted data sets the gamma GLM was applied to the data with an inverse link. The deviance for the gamma distribution is

$$2\sum_i \{-log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\},$$

summing over $i = 1, ..., n$ observations. The Poisson GLM was applied to binned data sets with an identity link. The deviance for this distribution is

$$2\sum_i \{y_i\ log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}.$$

Results

Table 3.1 shows the regression coefficients of all models. For a general feel with the actual numbers in the coefficients table, a quick fast-forward is made to Section 4. When simulating epidemics, the epidemic size ($N = 10000$) and the contact rate ($\beta = 0.1$) are known. When the sample size is large, the coefficient β_1 is approximately the contact rate β, and the sample size can be calculated with $1 - \frac{\beta_1}{\beta_2}$. Additionally, β_2 is approximately β_1/N. When analyzing actual epidemics, the original values of β and N are unknown and the comparison cannot be made. However, when the sample size is sufficiently large, one could assume that β_2 is some factor smaller than β_1.

It can be seen in **Table 3.1** that the regression coefficients of x and x^2 are significant at the $p < 0.001$-level for all fitted models. None of the intercepts were significant at the $p < 0.05$-level, save for model 2. The β_1 of models 5-8 are a factor 7 larger than those of models 1, 3 and 4: this corresponds to the expected multiplication with the bin size of 7 days. The values of β_1 for model 1, 3 and 4 are close to the originally simulated contact rate of 0.1, and the values of β_1 for models 5-8 are close to the simulated weekly contact

rate $\beta = 0.7$.
Table 3.2 shows fit statistics of the eight fitted models. At a glance the two gamma GLMs have comparable results for both null and residual deviance, although the null deviance for the gamma GLM applied to jittered data is slightly lower. The residual deviance is on its turn slightly lower for the gamma GLM applied to the default data. The values of the underreported data correspond to these observations. The deviance values of the models applied to binned data were in general lower than the values of the models applied to unbinned data: however, the analysis was performed on bin level, so there are naturally less data points and the deviance should be lower. The null deviance of the weighted gamma GLM was higher than the null deviance of the Poisson GLM, and the residual deviance of the weighted gamma GLM was lower than the Poisson GLM. For the underreported data the Poisson GLM seemed to perform best. Looking at the dispersion of the model, it can be seen that model 5 and 6 are overdispersed, indicating that the variability in the data is bigger than would be expected based on the model. Model 5 and 6 are the weighted gamma GLMs: it is possible that the overdispersion is caused by large weights corresponding to many observations in a bin.

Table 3.1. Regression coefficients of models 1-8, fit on an epidemic simulated with N=10000 and beta=0.1

	Intercept	SE	Pr(>|t|)	x	SE	Pr(>|t|)	xˆ2	SE	Pr(>|t|)
M1: default	-2.5e-02	6.3e-02	6.9e-01	1.0e-01	1.0e-03	0.0e+00	-1.0e-05	0.0e+00	0.0e+00
M2: +underrep	-3.3e-01	8.5e-02	1.3e-04	4.2e-01	8.2e-03	0.0e+00	-1.7e-04	0.0e+00	0.0e+00
M3: bin+jitter	-4.7e-02	4.7e-02	3.2e-01	9.9e-02	1.0e-03	0.0e+00	-1.0e-05	0.0e+00	0.0e+00
M4: +underrep	-5.6e-02	3.9e-02	1.5e-01	9.9e-02	2.0e-03	0.0e+00	-4.0e-05	0.0e+00	0.0e+00
M5: bin+weight	-4.4e-01	1.1e+00	6.8e-01	7.1e-01	3.0e-02	0.0e+00	-7.0e-05	0.0e+00	0.0e+00
M6: +underrep	-4.5e-01	6.8e-01	5.2e-01	7.1e-01	3.8e-02	0.0e+00	-2.8e-04	1.0e-05	0.0e+00
M7: bin+pois	3.2e-02	2.9e-01	9.1e-01	6.9e-01	7.0e-03	0.0e+00	-7.0e-05	0.0e+00	0.0e+00
M8: +underrep	-1.0e+00	6.4e-01	1.1e-01	6.8e-01	1.4e-02	0.0e+00	-2.8e-04	1.0e-05	0.0e+00

Table 3.2. Fit statistics (Deviance and dispersion) of models 1-8, for an epidemic simulated with N=10000 and beta=0.1

	Null deviance	df	Deviance	df	dispersion
M1: default	33508	9998	11552	9996	0.99
M2: +underrep	8539	2480	2777	2478	0.97
M3: bin+jitter	32690	9998	11842	9996	1.05
M4: +underrep	7897	2548	3019	2546	1.01
M5: bin+weight	19671	27	414	25	17.78
M6: +underrep	4723	25	148	23	6.94
M7: bin+pois	17966	25	517	23	1.00
M8: +underrep	3766	22	91	20	1.00

Model 1 and 2: gamma GLM for unbinned data

In model 1, an infection was simulated without binning, without underreporting and with incubation time. Subsequently, a gamma GLM was applied with an inverse link on an individual level.

It can be seen in **Table 3.1** that the coefficients for x and x^2 are significant at the $p < 0.01$-level. The coefficient for x is 0.1 and -0.00001 for x^2. The null deviance was 33508 with 9998 degrees of freedom. The residual deviance was 11552 with 9996 degrees of freedom: adding the covariates x and x^2 to the model improved the deviance with 21956 with 2 degrees of freedom: quite an improvement.

For model 2, an infection was simulated without binning, with a reporting rate of 25% and with incubation time. The coefficient for x is 0.42 and -0.00017 for x^2. This is quite large compared to the coefficients of model 1. The null deviance was 8539 with 2480 degrees of freedom. The residual deviance was 2777 with

2478 degrees of freedom. The residual deviance and corresponding degrees of freedom are much smaller due to the number of observations: from 10.000 to 2500. Adding the covariates x and x^2 to the model improved the deviance with 5762 with 2 degrees of freedom.

Model 3 and 4: gamma GLM for binned data: jittering binned data

In binned data there is little information about time until the next infection event. However, we can add some noise to the binned data and pretend that this noise is the information until the next event. With jittering, binned data can be stretched to data on the smallest unit, which is the individual level in our case. For each reported event, a random number from the uniform[0,1] distribution was drawn, multiplied by the time period Δt, and subsequently added to the binned time of reporting.

In model 3, an infection was simulated with binning, without underreporting and with incubation time. Subsequently, the data was jittered and a gamma GLM was applied with an inverse link. The coefficient for x is 0.1 and -0.00001 for x^2. The null deviance was 32690 with 9998 degrees of freedom. The residual deviance was 11842 with 9996 degrees of freedom: adding the covariates x and x^2 to the model improved the deviance with 20848 with 2 degrees of freedom.

Model 4 was the underreported version of model 3. The coefficient for x is 0.1 and -0.00004 for x^2. The null deviance was 7897 with 2548 degrees of freedom. The residual deviance was 3019 with 2546 degrees of freedom: adding the covariates x and x^2 to the model improved the deviance with 4878 with 2 degrees of freedom.

Model 5 and 6: Weighted gamma GLM for binned data

Applying a weighted gamma GLM was done on a week level: the predictor variable x was the cumulative number of infections; x^2 was the squared cumulative number of infections. The weights n were the number of observations in each bin; the response variable y was the average time to event, which was defined as $1/n$ in our case. **Table 3.3** and **Table 3.4** show a binned data set and a binned data set that was prepared for analysis with a weighted gamma GLM.

Table 3.3. Binned data set

x	x^2	y
3	9	3
9	81	6
34	1156	25
78	6084	44
256	65536	178

Table 3.4. Binned data set, prepared for weighted analysis

x	x^2	y	n
3	9	0.33	3
9	81	0.17	6
34	1156	0.04	25
78	6084	0.02	44
256	65536	0.01	178

In model 5, an infection was simulated with binning, without underreporting and with incubation time. The data was analyzed with a gamma GLM with an inverse link and weights for each bin corresponding to the number of events in that bin. The coefficient for x is 0.71 and -0.00007 for x^2. The null deviance was 19671

with 27 degrees of freedom. The residual deviance was 414 with 25 degrees of freedom: adding the covariates x and x^2 to the model improved the deviance with 19257 with 2 degrees of freedom. As Model 5 and 6 were analyzed on weekly basis, the coefficients should approximately be 7 times larger than the coefficients of Models 1 through 4. Except for Model 2 (underreported default) this assumption holds. Model 2 was fit on an underreported data set, so it is well possible that this model did not fit the underreported data properly. Applying a bootstrap would show us whether the high coefficients for Model 2 are random or whether this is a more structural result.

Model 6 was the underreported version of model 5. The coefficient for x was 0.71 and -0.00028 for x^2. The null deviance was 4723 with 25 degrees of freedom. The residual deviance was 148 with 23 degrees of freedom: adding the covariates x and x^2 to the model improved the deviance with 4575 with 2 degrees of freedom.

Model 7 and 8: Poisson GLM for binned data

The binned data consisted of event counts per time bin. A Poisson GLM with an identity link was applied to the data. In model 7 an infection was simulated with binning, without underreporting and with incubation time. The data was analyzed with a Poisson GLM with an identity link. The coefficient for x is 0.69 and -0.00007 for x^2. Again, in both model 7 and 8 the coefficients were larger with approximately a factor of 7. The null deviance was 17966 with 25 degrees of freedom. The residual deviance was 517 with 23 degrees of freedom. Adding the covariates x and x^2 to the model improved the deviance with 17449 with 2 degrees of freedom.

Model 8 was the underreported version of model 7. The coefficient for x is 0.68 and -0.00028 for x^2. The null deviance was 3766 with 22 degrees of freedom. The residual deviance was 91 with 20 degrees of freedom. Adding the covariates x and x^2 to the model improved the deviance with 3675 with 2 degrees of freedom.

It can be seen that the results for the binned approaches (Models 5 through 8) are very similar in both fully and underreported situations.

4. Calculating back Beta and N

We are interested in whether important information of an epidemic can be estimated back from the fitted regression models. When a model describes the epidemic well, we would expect that calculating back important quantities should be accurate. In this sense, the regression coefficients themselves are not so interesting: it is the information which can be obtained through them.

The infection rate β and the population size N were calculated from the regression coefficients. As the data was simulated according to a gamma distribution, we can use the distributional properties of the data set to approximate β and N.

In Section 2.1 we explained the simulation process. A short recap of the simulation:

- the number of infectious individuals at time t is $I(t)$;
- the population size is N;
- the contact rate β is the rate with which each infectious individual makes infectious contacts per day;
- the infection rate at time t is $\lambda(t)$;
- the time until next infection event, τ is given by an exponential distribution: $P(\tau|\lambda(t)) = \lambda(t)e^{\lambda(t)\tau}$;
- the expected time to the next event is $\boldsymbol{E}(\tau) = 1/\lambda(t)$.

Events are generated according to

$$\frac{1}{\boldsymbol{E}(\tau)} = \lambda(t) = \beta I(t)(1 - \frac{I(t)-1}{N-1}) = \beta\frac{N}{N-1}I(t) - \beta\frac{1}{N-1}I(t)^2$$

with τ distributed according to an exponential distribution.

The gamma regression is fit with an inverse link

$$\frac{1}{\boldsymbol{E}(\tau)} = \beta_0 + \beta_1 c + \beta_2 c^2$$

where τ is assumed to follow a gamma distribution.

When the event counter c is equal to the number of infectious individuals $I(t)$, the equations for generating events and fitting these events according to the gamma regression are exactly equal to each other when $\beta_0 = 0$. β and N can then be estimated from the regression coefficients as follows:

$$\beta = \beta_1\frac{N-1}{N}$$

$$N = 1 - \frac{\beta_1}{\beta_2}(1 - \frac{1}{N})$$

In the case of underreporting, the reporting rate p is the probability that the infection of a specific individual ends up in a surveillance system, and we assume that the individual probabilities are independent. In this case the event counter c can taken to be equal to the expected number of infectious individuals that are reported, $pI(t)$. We have

$$\beta \approx \beta_1$$

and

$$pN \approx 1 - \frac{\beta_1}{\beta_2}$$

This means that the contact rate can be inferred from the regression coefficients, regardless of underreporting, and that the population size N can be inferred from the regression coefficients if the reporting rate is known (approximately).

To illustrate the back-calculation of β and N, the results are used of the GLMs applied in Section 3. **Table 4.1** shows that the back-calculation for both β and N is quite accurate for this simulation: β is approximately 0.1 for the models applied on individual level and approximately 0.7 for the models applied on bin (week)

Table 4.1. Back-calculated values of beta and N

	Beta	N
Input	0.100	10000
M1: default	0.101	10076
M2: +underrep	0.415	2443
M3: bin+jitter	0.099	9856
M4: +underrep	0.099	2466
M5: bin+weight	0.706	10085
M6: +underrep	0.715	2553
M7: bin+pois	0.693	9895
M8: +underrep	0.681	2432

level. For respectively fully and underreported data the values approximate the original 10000 and 2500. The only outlier is the underreported default model, which consequently overestimated β after doing some checks.

The results in **Table 4.1** look very nice, however, the simulated data in this section might accidentally have approximated an ideal situation. In the next chapter a bootstrap will be performed to assess the predictive value of our approach and the consistency of these results. β and N are estimated as well, and the bias and variance of the epidemic parameters are calculated from the bootstrap samples.

5. Prediction

When infection data is analyzed, the parameters describing the epidemic are the most interesting. In the previous section regression coefficients were used to estimate the original epidemic parameters, β and N. We would like to know if the model that was used to create these estimates is somewhat reliable. This implies that the performance of the model should be checked in terms of model fit. Additionally, the forecasts following from this model should approach the original epidemic.

The following two sections are about prediction and forecasting. These are not to be confused: prediction is used *within* the range of observed values, to see whether a model is able to reproduce the original data when only a subset of the data is available for model fitting. The remaining data can then be used to create fit statistics from the predictions. Forecasting is used to extrapolate *beyond* the data range of available data. With forecasting the future course of the epidemic at hand is guessed. Forecasting intervals are used to show the most probable course of the epidemic. Because this is a simulation study, the cut-off point of the epidemic (for instance, 3 weeks) can be controlled and the originally simulated epidemic is available. This allows us to compare the forecasts to the original epidemic.

Predictions

A completed epidemic was simulated in the form of a time-to-event data set with the parameters $\beta = 0.1$ and $N = 1000$. β represents the contact rate (the average number of infectious contacts a person has during a day) and N is the population count. The data was divided into a training and a test set at the point $c_{halt} = 0.5$, which was when half of the population ($N = 500$) was infected during the original epidemic. The data before c_{halt} was the training set and the data after this point the test set. A GLM was fit on the training set, as in Section 3 on regression. With the obtained regression coefficients the values $\hat{\beta}$ and $\hat{N}$ were estimated as in Section 4. The values of $\hat{N}$ and $\hat{\beta}$ were used to peform a parametric bootstrap in which 500 new data (training) sets were created until the same cut-off point c_{halt}. We use $\hat{\beta}$ and $\hat{N}$ to simulate new training sets in stead of resampling the original training set. Again, GLMs were fit, and bootstrapped regression coefficients $\hat{\beta_0}^*, \hat{\beta_1}^*$, and $\hat{\beta_2}^*$ and bootstrapped estimates $\hat{\beta}^*$ and $\hat{N}^*$ were estimated. The bootstrap results were used to assess model fit, the regression coefficients, and the quality of the predictions made for the time to event τ by our method.

Similar to Section 3 eight different epidemics were simulated with variations on the reporting rate and binning. All epidemics were including incubation time. Three types of GLMs were fit to the corresponding epidemics. An overview of the simulations and fitted models can be seen in **Table 5.1**. The eight-model set-up was used again in this section for creating bootstrapped predictions. In short, every even-numbered epidemic (2, 4, 6, 8) was the underreported version of the preceding epidemic, with a reporting rate of 25%. Epidemic 1 was the default model, unbinned and fully reported. Epidemic 3 was binned and subsequently jittered data. gamma GLMs with an inverse link function were fit to these epidemics. Epidemics 5-8 were were analyzed on the level of binning, which was weeks in this study. Epidemics 5 and 6 were analyzed with weighted gamma GLMs with inversed link and epidemics 7 and 8 were analyzed with Poisson GLM with identity link.

The model fit was assessed with bootstrapped deviance and MSE statistics. Subsequently, bootstrapped fit statistics for the regression coefficients β_0, β_1 and β_2 are discussed, as well as for the estimated values of β and N.

Prediction accuracy: Deviance

For every bootstrap iteration, the null deviance, residual deviance and corresponding degrees of freedom were obtained. The results of the bootstrap can be seen in **Table 5.2** for all eight models.

The deviance is a measure of model fit. As described in Section 3 on regression, the smaller the deviance, the better the model fits the data. Two types of deviance were used to assess model fit: the null deviance and the residual deviance. The null deviance is the deviance between the saturated model (every data point has a

Table 5.1. Simulated epidemics and analysis methods used

Model	Description
M1: default	Incubation time, no binning, no underreporting, gamma GLM
M2: + underrep	Incubation time, no binning, underreporting, gamma GLM
M3: bin+jit	Incubation time, binning, no underreporting, jitter + gamma GLM
M4 + underrep	Incubation time, binning, underreporting, jitter + gamma GLM
M5: bin+weight	Incubation time, binning, no underreporting, weighted gamma GLM
M6 + underrep	Incubation time, binning, underreporting, weighted gamma GLM
M7: bin+pois	Incubation time, binning, no underreporting, Poisson GLM
M8: + underrep	Incubation time, binning, underreporting, Poisson GLM

parameter) and the null model (intercept only). The residual deviance is the deviance between the model we want to fit (systematic model of the GLM) and the saturated model.

To test whether the proposed (systematic) model is an improvement of the null model (only an intercept), one could perform a χ^2-test on the difference in deviance with the difference in degrees of freedom between these models. A p-value below 0.05 would indicate that the proposed model is an improvement of the null model. This test was performed and can be seen in the 'null.dev' rows.

When a χ^2-test is performed on the residual deviance or the null deviance itself with its corresponding degrees of freedom, in essence it is tested whether the model can be improved in the direction of the saturated model: in short, it is tested whether adding parameters would improve model fit. We do not show the null deviance test, as this is not interesting; we do show the residual deviance test to see whether our systematic model can be improved.

To see whether the chosen systematic model ($\beta_0 + \beta_1 x + \beta_2 x^2$) can be improved further, the residual deviance would be the χ^2-value to be tested on its corresponding degrees of freedom. A p-value below 0.05 would indicate that the model can still be improved. The 'deviance' rows of the 'p-value' column of **Table 5.2** show the p-values for this test.

For example for the default model, the bootstrapped mean residual deviance is 574 on 497 degrees of freedom. A $\chi^2(574, 497)$-test was performed, yielding a p-value of 0.01. This indicated that this model could be improved.

Looking at **Table 5.2**, several things stand out: first, the residual deviance statistics in Models 5-8 are much smaller than the residual deviance statistics of Models 1-4. We assign the difference mainly to the level of analysis for the predictor variables: Models 1-4 are analyzed on an individual level, and Models 5-8 on a weekly level. The number of parameters in Models 5-8 is much smaller in the latter models. The deviance of model 5 and 7 remains at roughly zero: this does not mean that the model is perfect, rather that the weighted gamma GLM and the Poisson GLM should probably not be measured with the deviance: there is much less variation in binned data. This would also explain the very low deviance statistics for all binned models that were analyzed on a week level.

Mean Squared Error for model fit

The MSE measures the difference between the true value of some statistic and its estimator. Additionally, it can be used to measure the closeness of the predicted data to the original data. Here, the MSE was calculated as the mean of the squared difference between predictions of the time to event variable $\hat{\tau}$, and the original observations of τ:

$$MSE = E[(\hat{\tau} - \tau)^2].$$

The MSE was calculated for all eight simulated epidemics for 500 bootstraps. **Table 5.3** shows the mean, median, minimum, maximum, 95% bootstrap confidence intervals and square root of the MSE.

Table 5.2. Residual Deviance and Null Deviance

	mean	sd	median	5%	95%	min	max	df	chisqtest
M1: deviance	574	34	575	519	630	464	680	497	0.01
null.dev	1236	171	1215	1008	1543	922	2007	499	0.00
M2: deviance	79	15	78	57	107	40	129	65	0.11
null.dev	152	46	147	92	242	67	353	67	0.00
M3: deviance	591	36	589	534	656	486	703	497	0.00
null.dev	1132	127	1122	945	1374	862	1662	499	0.00
M4: deviance	36	11	35	20	54	9	77	27	0.12
null.dev	91	27	89	51	139	27	190	29	0.00
M5: deviance	0	0	0	0	0	0	2	0	1.00
null.dev	385	189	398	47	714	0	904	1	0.00
M6: deviance	8	12	0	0	33	0	73	1	0.01
null.dev	50	20	49	23	86	0	114	3	0.00
M7: deviance	0	0	0	0	0	0	0	0	1.00
null.dev	379	196	396	33	726	0	950	2	0.00
M8: deviance	6	12	0	0	31	0	79	1	0.02
null.dev	45	19	44	13	77	0	119	3	0.00

In general it can be seen that the distribution of the bootstrapped MSE values is skewed: the median MSE is substantially lower than the mean MSE for all models. This indicates that there are several epidemics which are very different from the original epidemic, but that the majority is not extremely different. For example in the default model, it can be seen that the range $(0.64, 22473)$ of the MSE is quite broad over all bootstraps. The bootstrap mean of the MSE is 51.36. However, the median has a value of 0.65 showing that most predictions are quite close to the original data. The 95% confidence interval corresponds to the median with (0.64, 1.9). This implies that the majority of the predictions performs well, but we should be wary of the few that do not do so.

What also stands out are the 'inflated' MSE values of the Poisson GLM, and the MSE values of the weighted gamma GLM being between zero and one. Thinking about how this model is made, it is actually quite logical: the time to event for the Weighted gamma GLM is determined as 1 divided by the number of observations in a bin, thus it can range between 0 and 1 for either the predictions or original data. The difference between a predicted and an original time to event value can not become larger than 1, and the mean squared error is bound between 0 and 1. In this case, a MSE close to 1 indicates that the mean difference between the estimated number of observations in a bin, and the actual number of observations in this bin, is large. The MSE of the Poisson GLM is relatively large compared to the other GLM approaches. However, it should not be forgotten that the response variable in Poisson GLM consists of counts, and that a squared difference in counts can become quite large very quickly. Therefore, it is safe to assume that the MSE is proportional to the size of the original epidemic for Poisson GLM: a 10% difference creates a larger error if the count in a specific week is 500 (50^2) than if this count is 50 (5^2). To look at actual values one can take the square root of the MSE. For the results in **Table 5.3** there is a difference of 12 in $\sqrt{MSE}$ between the bootstrap mean MSE and median MSE for the Poisson GLM(which is not extreme), and the maximum $\sqrt{MSE}$ amounts to a a total error of 168 in the course of an epidemic (which is quite a lot).

Looking at the bootstrap median MSE of the gamma GLMs it can be seen that the gamma GLM applied to jittered data performs slightly worse for the fully reported data and substantially worse for underreported data. Taking the root of the median MSE values, the gamma GLM applied to jittered, underreported data amounts to an average discrepancy of 2.8 days for an observation, whereas the gamma GLM applied to original data amounts to a discrepancy of 0.2 days. The gamma GLMs applied to fully reported data are closer in terms of root MSE: the average discrepancy is approximately 0.8 days for the gamma GLM applied to the default data, whereas this is 1.1 days for the gamma GLM applied to jittered data.

Table 5.3. Mean Squared Errrors

	mean	sd	median	5%	95%	min	max	sqrt(MSE)
M1: default	51.36	858.43	0.65	0.64	1.90	0.64	22473	0.81
M2: + underrep	0.10	0.28	0.05	0.05	0.14	0.05	2	0.22
M3: bin+jit	3.35	26.30	1.30	1.29	1.85	1.29	648	1.14
M4 + underrep	15.90	54.30	7.64	7.57	14.05	7.57	860	2.76
M5: bin+weight	0.70	0.39	0.92	0.10	1.08	0.10	1	0.96
M6 + underrep	0.12	0.12	0.07	0.04	0.29	0.04	1	0.26
M7: bin+pois	7120.22	5551.14	5492.39	3269.26	22453.57	3265.78	28547	74.11
M8: + underrep	552.11	471.41	253.78	173.78	1541.78	172.79	3782	15.93

Fit statistics of the regression coefficients

Fit statistics for the regression coefficients were obtained as well. **Table 5.4, 5.5** and **5.6** show the mean, standard deviation, median, 5% and 95% bootstrap confidence interval of the regression coefficients β_0, β_1, and β_2, as well as the number of NA values.

Save model 6, none of the intercepts are significantly different from zero, based on the 90% bootstrap confidence intervals given in **Table 5.4**. This implies that the epidemic starts with zero infected individuals. The range, variance and MSE of the intercepts are very large for models 5-8: there might be some issues with the models applied on binned data. The bias estimates for fully reported weighted gamma GLM and Poisson GLM seem to be acceptable, but this is a trade-off with the extremely large variance. Overall for the intercept, the models applied on an individual level perform best when applied to fully reported models.

According to the 90% bootstrap confidence intervals, all estimates of the regression coefficient β_1 are significantly different from zero. The estimates of β_1 for models 5-8 applied on bin level have a less extreme range than the intercept, as can be seen in **Table 5.5**. The maximum values of β_1 for models 5-8 range from 96 for the weighted gamma GLM applied to underreported data, to 514 for Poisson GLM applied to fully reported data. The range of the models applied on an individual level is acceptable, although the gamma GLM applied to underreported default data had a relatively high maximum value (2.6) and a higher range (0.9;1.98) than the other GLMs applied on individual level.

β_1 should approximate the epidemic parameter $\beta = 0.1$ for models applied on individual level with a multiplication of bin size for analyses on bin level. Only models 1 and 3 in which the gamma GLM is applied to individual level data approximately reproduce the original epidemic parameter. Additionally, the range and fit statistics of these two models look promising. However, the bootstrapped means of β_1 for model 2 (0.13), model 4 (-0.08), and models 5-8 (repectively 0.3, -0.11, 0.82 and 0.1) are larger than the expected value of $\beta_1 \approx 0.7$ given the multiplication of 7 days. The bias, variance and MSE of models 2 and 4 are not so dramatic; the bias, variance and MSE of models 5-8 are substantial.

As calculated in the previous section, and as mentioned in section 3 on regression, β_2 is approximately a factor N smaller than β_1. It can be seen in **Table 5.6** however, the multiplication factor is a bit below a factor of 100 in our case. The estimated of models 4, 5, and 7 are not significant according to the 90% bootstrap confidence interval. Again, the bias, variance and MSE of the models applied to binned data are inflated compared to the fit statistics of models applied to individual level data.

The number of not estimated regression coefficients (the 'NAs' columns in **Table 5.4**, **5.5** and **5.6**) was not discussed before, as there were no NAs in the first two regression coefficients β_0 and β_1 for any model. However, for β_2 there are not-estimated regression coefficients for the binned models: 229 for model 5, 8 for model 6, 214 for model 7, and 28 for model 8. Actually, when introducing the quadratic term in our systematic model, an exact linear relation between x and x^2 was defined. The consequence of this exact collinearity is that there is no unique solution to the regression, and the GLM algorithm drops the x^2 variable to estimate the model with only the variables that contribute to the regression. This artificially introduced

Table 5.4. Bootstrap statistics for beta 0

	mean	sd	median	5%	95%	min	max	bias	variance	MSE	NAs
M1: default	0.06	0.13	0.01	-0.08	0.30	-0.11	0.82	0.10	0.02	0.03	0
M2: + underrep	0.85	3.16	-0.27	-1.40	7.24	-2.33	21.89	0.57	9.96	10.28	0
M3: bin+jit	-0.03	0.07	-0.04	-0.10	0.11	-0.13	0.29	0.07	0.00	0.01	0
M4 + underrep	-0.22	0.24	-0.15	-0.70	0.06	-1.04	0.24	-0.11	0.06	0.07	0
M5: bin+weight	1.15	202.73	0.00	-18.32	95.60	-4112.80	680.15	1.04	41015.43	41016.51	0
M6 + underrep	-144.52	322.36	-27.24	-776.86	-0.85	-2241.78	56.89	-144.93	103705.66	124710.65	0
M7: bin+pois	1.14	347.40	-0.05	-18.83	123.25	-7468.62	1053.08	1.21	120445.88	120447.35	0
M8: + underrep	-152.10	317.59	-28.21	-659.38	7.76	-2550.93	257.14	-152.41	100664.01	123892.06	0

Table 5.5. Bootstrap statistics for beta 1

	mean	sd	median	5%	95%	min	max	bias	variance	MSE	NAs
M1: default	0.11	0.01	0.11	0.09	0.12	0.08	0.14	0.00	0.00	0.00	0
M2: + underrep	1.39	0.33	1.36	0.90	1.98	0.43	2.60	1.02	0.11	1.16	0
M3: bin+jit	0.13	0.01	0.13	0.12	0.15	0.10	0.16	0.00	0.00	0.00	0
M4 + underrep	0.29	0.22	0.20	0.07	0.74	-0.03	1.15	0.15	0.05	0.07	0
M5: bin+weight	2.14	8.75	1.00	0.46	4.64	-1.03	173.79	1.48	76.48	78.66	0
M6 + underrep	11.63	17.38	5.55	1.70	41.74	-0.78	95.98	11.12	301.39	424.96	0
M7: bin+pois	3.05	23.33	1.05	0.37	5.84	-1.85	513.88	2.36	543.23	548.78	0
M8: + underrep	12.72	17.59	6.31	0.56	38.60	-3.29	107.39	12.18	308.95	457.41	0

collinearity poses an issue when one wants to estimate $\hat{\beta}$ and $\hat{N}$ from the regression coefficients.

Fit statistics of β and N

The bias, variance and MSE, 90% bootstrap confidence intervals, range, and NA values of bootstrapped estimates $\hat{\beta}^*$ and $\hat{N}^*$ were assessed. The aim was to see how close the estimated values were to the original values of the contact rate β and the population size N. The results can be seen in **Table 5.7** and **5.8**.

The bias of the bootstrap estimator $\hat{N}^*$ of original parameter N is defined as

$$Bias(\hat{N}^*, N) = E_N[\hat{N}^*] - N = E_N[\hat{N}^* - N].$$

The variance is defined as

$$Var(\hat{N}^*) = E[(\hat{N}^* - E[\hat{N}^*])^2].$$

Table 5.6. Bootstrap statistics for beta 2

	mean	sd	median	5%	95%	min	max	bias	variance	MSE	NAs
M1: default	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
M2: + underrep	-0.02	0.01	-0.02	-0.03	-0.01	-0.04	-0.01	-0.02	0.00	0.00	0
M3: bin+jit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
M4 + underrep	-0.01	0.01	0.00	-0.02	0.00	-0.05	0.01	0.00	0.00	0.00	0
M5: bin+weight	-0.01	0.02	0.00	-0.02	0.00	-0.33	0.00	0.00	0.00	0.00	229
M6 + underrep	-0.15	0.20	-0.08	-0.47	-0.02	-0.97	0.00	-0.15	0.04	0.06	8
M7: bin+pois	-0.01	0.06	0.00	-0.02	0.00	-1.00	0.00	-0.01	0.00	0.00	214
M8: + underrep	-0.18	0.22	-0.10	-0.60	-0.02	-0.97	0.00	-0.18	0.05	0.08	28

Table 5.7. Bootstrap fit Statistics for Beta

	mean	sd	median	5%	95%	min	max	bias	variance	MSE
M1: default	0.11	0.01	0.11	0.09	0.12	0.08	0	0.01	0.00	0.00
M2: + underrep	1.39	0.33	1.36	0.90	1.98	0.43	3	1.29	0.11	1.77
M3: bin+jit	0.13	0.01	0.13	0.12	0.15	0.10	0	0.03	0.00	0.00
M4 + underrep	0.29	0.22	0.20	0.07	0.74	-0.03	1	0.19	0.05	0.08
M5: bin+weight	2.13	8.75	1.00	0.46	4.63	-1.03	174	2.03	76.33	80.46
M6 + underrep	11.62	17.36	5.55	1.70	41.70	-0.78	96	11.52	300.78	433.54
M7: bin+pois	3.05	23.31	1.05	0.36	5.83	-1.84	513	2.95	542.14	550.85
M8: + underrep	12.70	17.58	6.30	0.56	38.56	-3.28	107	12.60	308.33	467.19

Table 5.8. Bootstrap fit Statistics for N

	mean	sd	median	5%	95%	min	max	bias	variance	MSE
M1: default	881	134	855	724	1128	673	2062	-119	17969	32087
M2: + underrep	72	9	71	59	86	34	106	-928	78	861814
M3: bin+jit	541	19	536	518	583	512	625	-459	379	211118
M4 + underrep	-175	4952	44	-224	223	-110240	5326	-1175	24472256	25853290
M5: bin+weight	4038	43273	881	519	4456	510	713380	3038	1865603850	1874830568
M6 + underrep	74	16	72	59	91	49	357	-926	258	857277
M7: bin+pois	1322	2220	777	515	3470	505	29067	322	4912229	5016105
M8: + underrep	73	16	71	56	91	45	282	-927	255	859204

the Mean Squared Error is defined as:

$$MSE(\hat{N}^*) = E[(\hat{N}^* - N)^2],$$

where N represents the true population size and $\hat{N}^*$ the bootstrap estimates of N. We can decompose it with the bias-variance decomposition into

$$MSE(\hat{N}^*) = Var(\hat{N}^*) + [E(\hat{N}^*) - N]^2.$$

For β, one can simply apply the same formulas, substituting N for β.

The similarity of **Table 5.4** and **Table 5.7** is striking. As briefly mentioned in Section 3, and as shown in Section 4, when N is sufficiently large, β_1 should approximate the original value of $\beta = 0.1$. In this case, the bootstrapped estimates of β_1 are almost identical to the bootstrapped estimates of $\hat{\beta}$.

The bootstrapped statistics for $\hat{N}$ in **Table 5.8** show that there were some issues: for several models very extreme values were estimated: the gamma GLM applied to underreported data had a minimum value of -110240; maximum values of the weighted gamma GLM and Poisson GLM applied to fully reported data were respectively 713380 and 29067. It can be seen that the presence of extreme values has inflated the MSE and variance of these models. Additionally, the means of these three models were influenced and for interpretation the median was used. The 90% bootstrap confidence intervals were less extreme, and the intervals of models 1, 5, and 7 contained the original value $N = 1000$.

It can be seen that the median of all underreported models amounts to estimates between 44 and 881, which is more than factor 10 smaller than the original value of $N = 1000$. The medians of the GLMs applied to fully reported data also underestimate N. This is reflected in the bias, which is negative for all models, save the weighted gamma GLM and Poisson GLM, which had a few extremely large estimates of N.

Conclusion

A simulation study of 500 bootstraps was performed to assess the fit statistics of several GLMs, of their coefficients and of the estimated values $\hat{\beta}$ and $\hat{N}$ from these coefficients. In general, the models applied to individual level data performed best out of the eight models applied to the data. Only the gamma GLM applied to individual level, fully reported data, and the gamma GLM applied to individual level, jittered data returned estimates of β_1 that corresponded to the original value of $\beta = 0.1$ used for simulating the data. The values of N have a very broad range, and only the 90% bootstrap confidence intervals of the gamma GLM applied to fully reported default data, the weighted gamma GLM applied to fully reported data, and the Poisson GLM applied to fully reported data, contained the original value $N = 1000$.

Regarding the fit statistics for the models, the MSE corresponds to the analysis level of the response variable used in the regression. This implies that it is not a useful measure for comparing models; however, it is a useful measure to assess the actual mean error for an observation of the response variable in some cases.

For models applied on bin level, one of the major issues arising is the exact collinearity introduced by the systematic model. In almost 50% of the bootstrap runs the coefficients of β_2 could not be estimated, because the variable was omitted from the model.

6. Forecasts

During an epidemic, it is necessary to extrapolate beyond the range of the data at hand and look into future time points. Forecasting allows us to make an educated guess at the future. A model was fit on the training data of Section 5 and forecasts were made from the end of the training data (at c_{halt}) until the end of the epidemic. Prediction intervals were created from the bootstrapped forecasts and compared to the original epidemic: as this is a simulation study, the original is available.

There are several approaches to making forecasts from the data. The first one is making forecasts exactly the same way the simulation was made: that implies when there was binning, again the forecasts will be binned. A second approach makes a default forecast (unbinned, not underreported but with incubation time) regardless of the original simulation. An advantage of the first approach is the easy comparability to the originally simulated epidemics. Another advantage is the realism introduced by not having the whole data set available. The disadvantage is that the information is reduced from specific time points to binned counts, possibly making the forecasts less accurate. The simulation in this project first creates an unbinned epidemic and bins it afterwards. As such, it can be argued that this third approach is still using as much information as the second approach, while preserving the "what you see is what you forecast" comparability to the originally simulated epidemics.

The procedure for forecasting is similar to that of prediction; however, we use the $\hat{\beta}^*$s and $\hat{N}^*$s from the bootstraps to create new forecasts instead of predictions. For the sake of completeness, the procedure is described again:

A completed epidemic was simulated in the form of a time-to-event data set with the parameters $\beta = 0.1$ and $N = 1000$. β represents the contact rate (the average number of infectious contacts a person has during a day) and N is the population count. The data was divided into a training and a test set at the point $c_{halt} = 0.5$, which was when half of the population ($N = 500$) was infected during the original epidemic. The time point corresponding to this cut-off point was variable, as can be seen in the forecasting figures that are shown later in this section. The data before c_{halt} was the training set and the data after this point the test set. A GLM was fit on the training set, as in Section 3 on regression. With the obtained regression coefficients the values $\hat{\beta}$ and $\hat{N}$ were estimated as in Section 4.
The values of $\hat{N}$ and $\hat{\beta}$ were used to peform a parametric bootstrap in which 500 new data sets were created until the same cut-off point c_{halt}. Again, GLMs were fit, and bootstrapped regression coefficients $\hat{\beta_0}^*, \hat{\beta_1}^*$, and $\hat{\beta_2}^*$ and bootstrapped estimates $\hat{\beta}^*$ and $\hat{N}^*$ were estimated. The bootstrapped values $\hat{\beta}^*$ and $\hat{N}^*$ were used to create forecasts from c_{halt} onward until the end of the epidemic (which is when all individuals are infected). The end point (in days) differs for each bootstrapped forecast. The bootstrapped epidemics from c_{halt} onward were used to make forecasting intervals for the cumulative number of infected individuals. Incubation time was included in all models.

It was assumed that 500 bootstrap samples were enough to create a daily interval of $\hat{N}^*$ by means of the R function `quantile()`. The `quantile()` function shows how much of the data at hand lies below a certain value. The 5% and 95% quantiles were picked to determine a daily 90% forecasting interval. The chosen interval was a bit more narrow than the usual 95% interval in statistics: a slightly more narrow forecasting interval might be a safeguard against outliers. The 50% quantile was used to get the median of the predictions. As the outliers in the estimated $\hat{N}^*$ could jump to extremely high numbers, elevating the mean, the median was a more appropriate statistic.

Important to note is that, to save computing time, the forecasts were *not* made when the estimated epidemic size, $\hat{N}^*$, was more than 25 times the original $N = 1000$ used for simulating the original epidemic. This bootstrap iteration was simply skipped. Additionally, sometimes the estimated value of the rate, $\hat{\beta}^*$, was zero or negative, or $\hat{N}^*$ was `NA`. These runs were also skipped as they would have impossible starting values. Skipping the iterations was chosen rather than picking a likely value (e.g. $N = 1000$) as this information is usually not available when using non-simulated data. Later in this section, the 'skipped' column in **Table 6.1** will be discussed, which shows the number of skipped bootstrap samples for each model.

Table 6.1. Median forecast and 90 percent forecasting intervals of number of cases, and duration of simulated epidemics. The actual value of N in the simulations was 1000. The actual duration of the whole epidemic was 150 days.

	N			duration			
	median	5%-ile	95%-ile	min	90%-ile	max	skipped
M1: default	855	725	1128	114	239	2885	0
M2: + underrep	534	520	553	60	65	95	0
M3: bin + jit	536	518	583	3	11	18	0
M4: + underrep	527	505	876	2	18	37	3
M5: bin + weight	880	519	4452	3	14	30	230
M6: + underrep	546	530	567	4	10	11	8
M7: bin+pois	774	515	3327	2	14	19	215
M8: + underrep	538	521	562	3	9	13	28

Results

Forecasts were made for models 1-8. First, the figures and tables are presented and briefly described. Subsequently the separate models will be discussed in more detail. The forecasting results are concluded and an overview is given of the issues that have arisen while making the forecasts.

In **Figure 6.1 to 6.4**, the forecasts were plotted for every bootstrap sample of Models 1-8, together with the forecasting intervals (red lines), the median (blue line) and the original epidemic (green line). On the left panels (indicated by 'a') are the full epidemics; on the right panels ('b') a zoomed in plot until the 90th quantile of the epidemic length in days (or in weeks if binned) of the bootstrapped simulations can be seen. The zoomed-in plot is shown, because sometimes an outlier will stretch the duration of the epidemic, having the last infection of the epidemic much later in time than the other simulations. It will seem (as in **Figure 6.1a**) as if the epidemics increase very quickly, and as if they take a very long time until completion; all epidemics will become horizontal and non-increasing for most of the plot. The zoomed-in plots can show more information on the actual epidemic forecasts than the stretched one.

Table 6.1 shows the forecasting intervals for models 1-8 for the number of infections when all forecasted epidemics were completed. It also shows the range and 90 percentile cut-off for the time the forecasts took from c_{halt} onward until completion, in days (model 1 and 2) or weeks (models 3-8). Models 3 and 4 were of jittered data: however, the original data was binned and the forecasts are subsequently binned as well. Note that the x-axes of the plots show from c_{halt} until epidemic end, taking into account the time before c_{halt} in the axis. In **Table 6.1** the duration in time from c_{halt} does not take into account the time before this point, to show the remaining time the epidemic would take.

The median estimates of the epidemic size were below the original $N = 1000$ for all eight models, ranging from 527 to 880. It can be seen in the 90% forecasting intervals that the gamma GLM applied to default data performs best in terms that it is closest to the original epidemic. Model 2, 3, 4, 6, and 8 systematically underreport the number of infections with about 450-480 compared to the the original epidemic. As models 2, 4, 6, and 8 were fit on 25% of the data, it was to be expected that they would underreport.

The estimated infection size for the gamma GLM applied to jttered data was different from the other models: the fully reported estimates $\hat{N}^*$ were lower than the other fully reported estimates. The gamma GLM applied to jittered epidemics was expected to perform approximately the same as the default model, due to the random variation added. It might be that the variation in the original epidemic follows a specific pattern that was missed when a uniform jitter was added. It would be possible to add specific jitter patterns when information is available on the distribution in a bin. For example, most practices close during weekends. That could lead to a bathtub shape each week in the reporting, with many reports on Monday and Friday. Or everyone would try to go to the doctor on Monday, creating a pattern where disease reporting peaks on Mondays and then decreases througout the week.

Regarding the duration of the epidemics the 90% quantile shows that most epidemics end within 15 weeks of the cut-off point of 50% infected. The gamma GLM is an outlier, as the 90th quantile is at 239 days (34 weeks). The original end point of the epidemic was approximately 150 days (in the 22th week). In the next subsections, figures of the forecasts will show more information on forcast durations.

Model 1: gamma GLM aplied to individual level data

It can be seen in **Figure 6.1a** that the majority of forecasts made with the default simulation was close to the original epidemic. The forecasting interval contained the original value of $N = 1000$, although the median was lower, showing that most forecasts underreported the original value. **Table 6.1** shows that the median was 145 off the original. It can be seen in **Figure 6.1a** and **b** that the duration of the original epidemic was shorter than all forecasted epidemics (take note of the varying x-axis). The original epidemic finished at approximately 150 days, however the value of c_{halt}, derived from the first estimate, was beyond that point. The black marks show that the majority of the forecasts ended between 200 and 300 days from the cut-off point.

Model 2: gamma GLM aplied to underreported, individual level data

It can be seen in **Figure 6.1c** that the forecasts made with the underreported default simulation underreported the original value of $N = 1000$. The forecasting interval (520 ; 553) did not contain the original epidemic, and the median was far below the original value with a difference of 466. The duration of the original epidemic was longer than most of the simulated epidemics, however, the durations of the forecasts made with the underreported data were closer to the original duration than the forecasts made with fully reported data.

Model 3: gamma GLM applied to jittered data

In **Figure 6.2a** and **b** it can be seen that the forecasts made with binned and jittered analysis underestimated the original epidemic size. Mind that these plots were made in weeks, the binning size, and that 20 weeks corresponds to 140 days. The forecasting interval (518 ; 583) did not contain the original epidemic, and the median was far from the original $N = 1000$. The green line of the original epidemic can be seen at the left of both plots, showing a rapid increase in week 13 (the starting point of these forecasts), whereas the forecasts themselves are increasing with only tens of infections per week. The duration of the original epidemic (about 22 weeks) was close to the 90th quantile of epidemic duration, although it can be seen that the majority of the forecasts ends between the 14th and 21st week.

Model 4: gamma GLM applied to underreported, jittered data

Figure 6.2c and **d** show that the forecasting interval (505 ; 876) made with the gamma GLM applied to underreported, jittered data neared the original epidemic size, although the median of 527 was just as far from the original epidemic as the other underreported forecasts. It can be seen that the underreported model produced a few very high estimates $\hat{N}^*$. For both model 3 and 4 it is interesting to see that the rapid growth that we expected from c_{halt} onward, starting the forecast halfway through the epidemic, did not show at all. Rather, it seems as if the epidemic started over at one infected individual again. An explanation would be that the model did not capture the exponential growth, passing on a small estimated infection rate $\hat{\beta}$. The duration of most forecasts was longer than the original epidemic. Although the maximum duration was 45 weeks from the start of the epidemic until epidemic end, the majority was beneath 27 weeks (189 days). The y-axis of these plots has a broader range than that of panels a. and b., however it is clear that the slow growth is visible here as well before week 20. The outliers in epidemic size are also in to 90+ quantiles of duration.

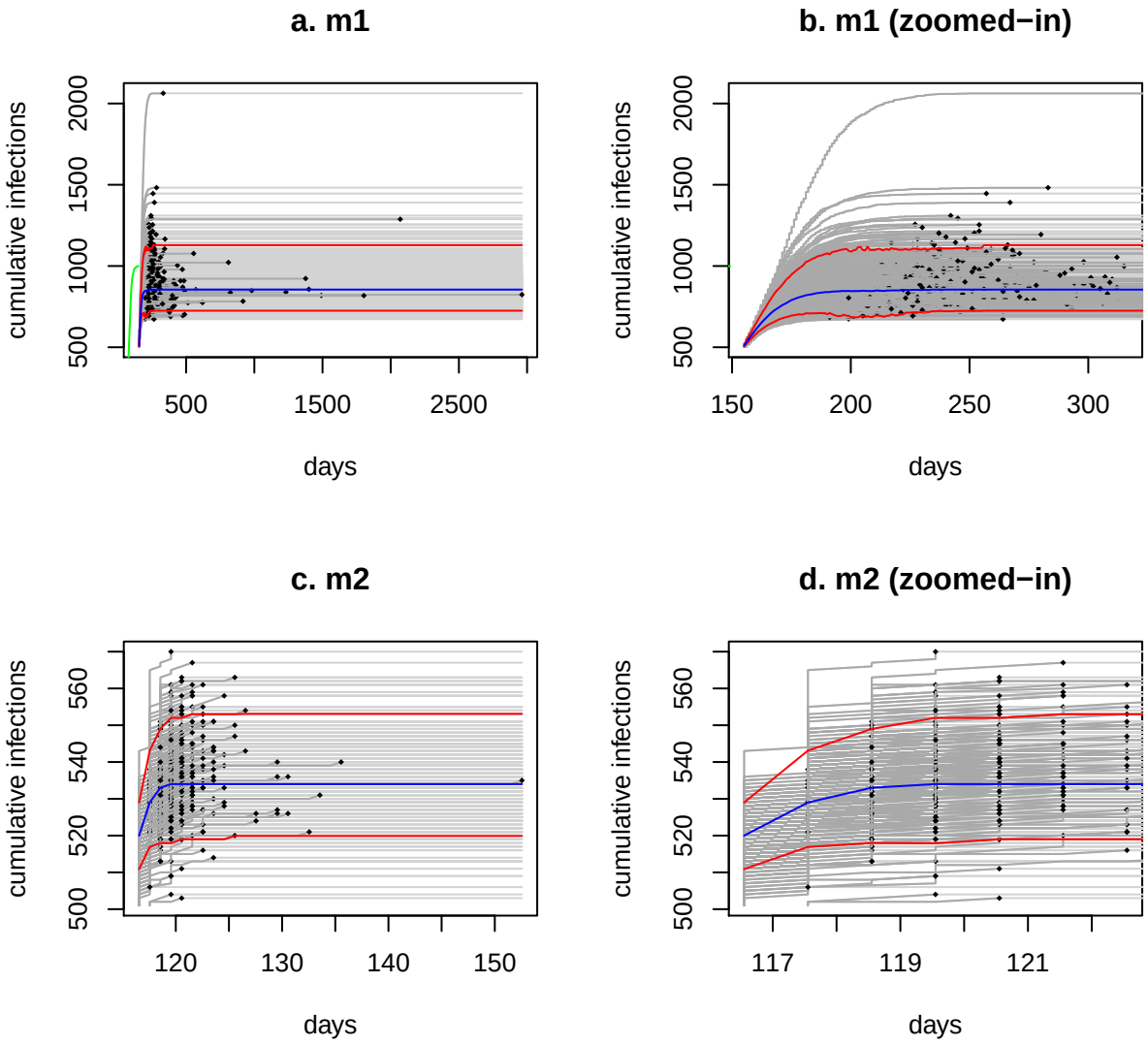


Figure 6.1. Forecasts for model 1 and 2: gamma GLM applied to individual level data. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panels (a and c) show the full length of the forecasts, whereas the right panels (panels b and d) show the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.

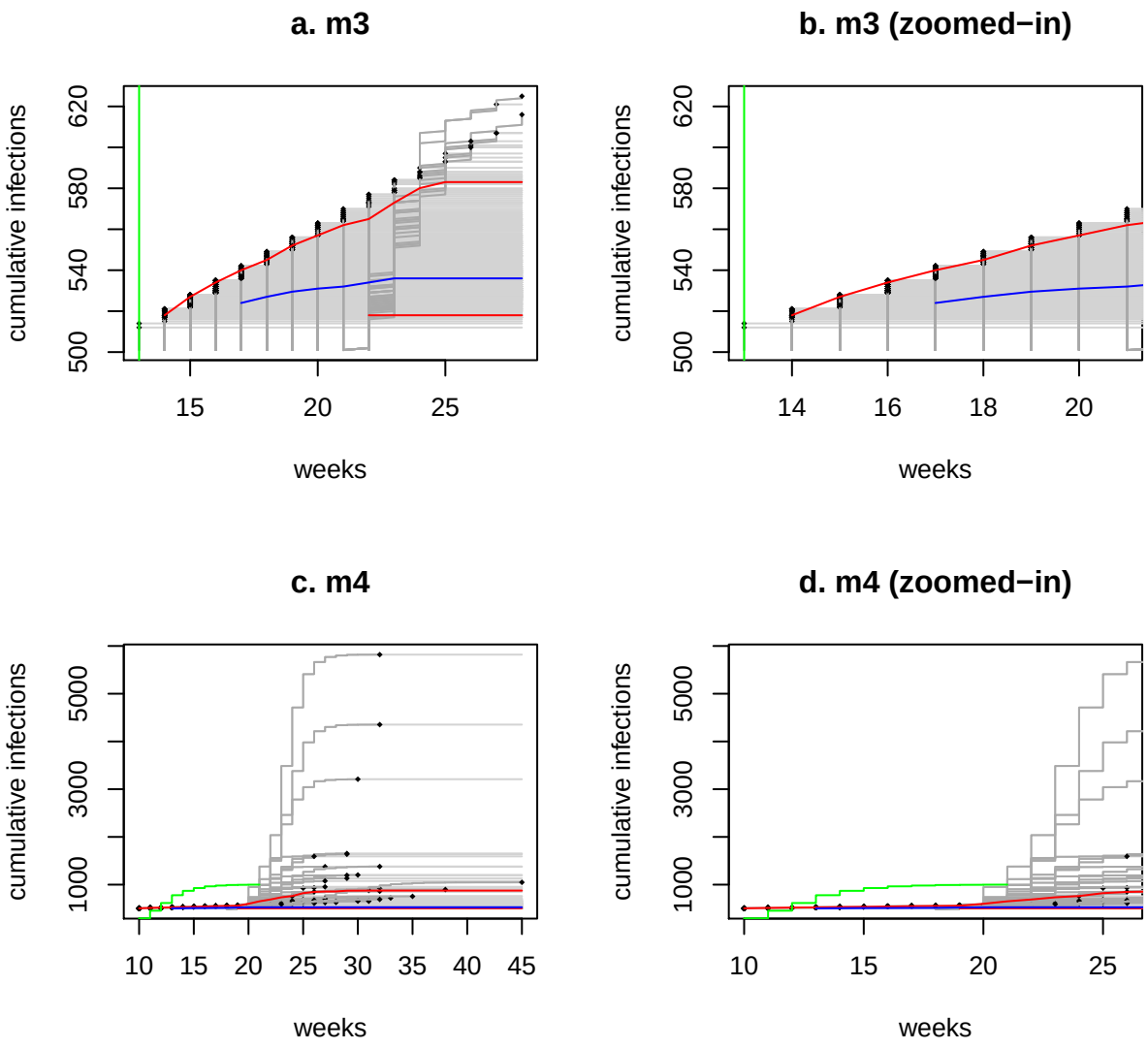


Figure 6.2. Forecasts for model 3 and 4: gamma GLM applied to individual level, jittered data. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panels (a and c) show the full length of the forecasts, whereas the right panels (panels b and d) show the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.

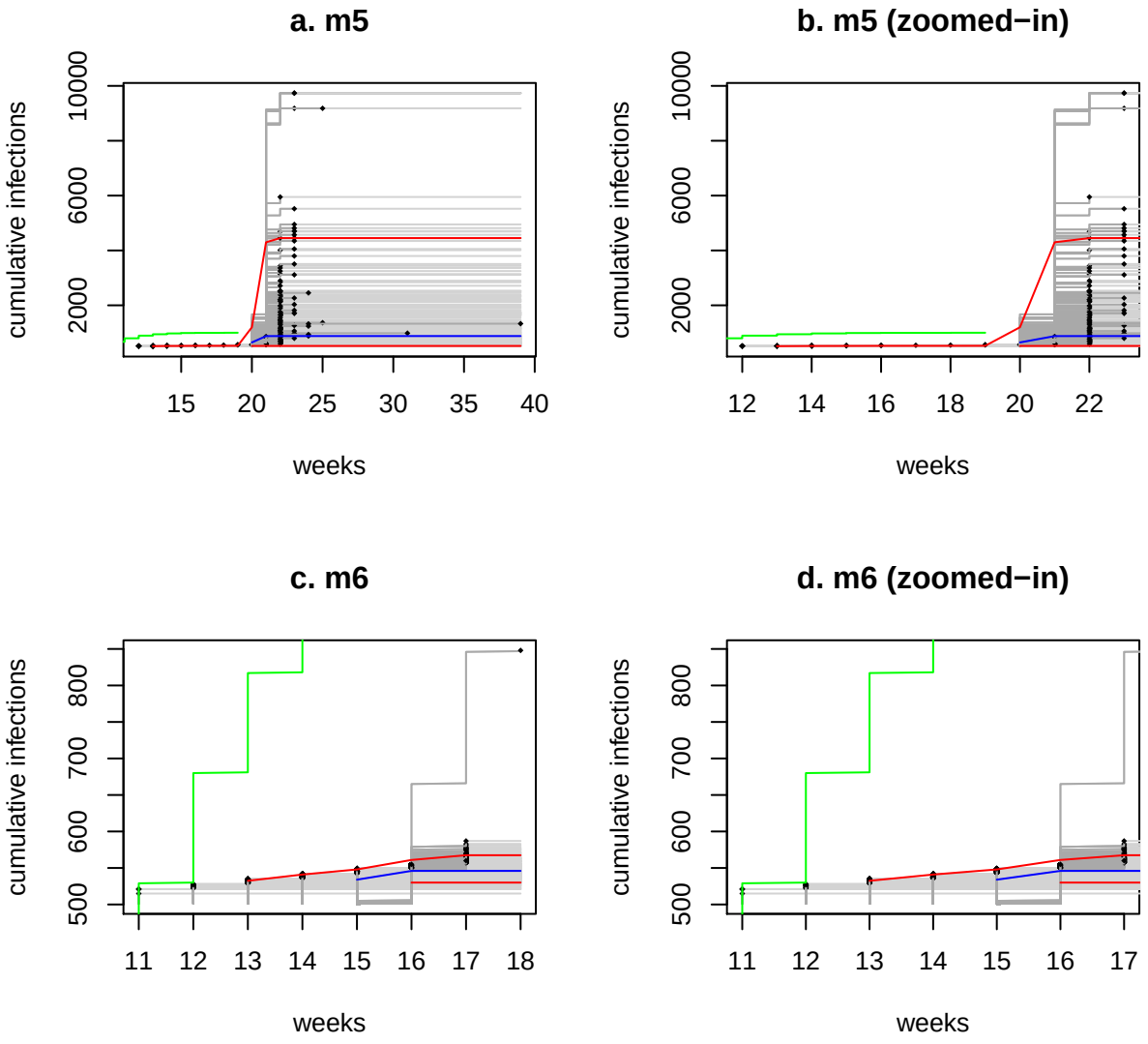


Figure 6.3. Forecasts for model 5 and 6: weighted gamma GLM applied to binned data. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panels (a and c) show the full length of the forecasts, whereas the right panels (panels b and d) show the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.

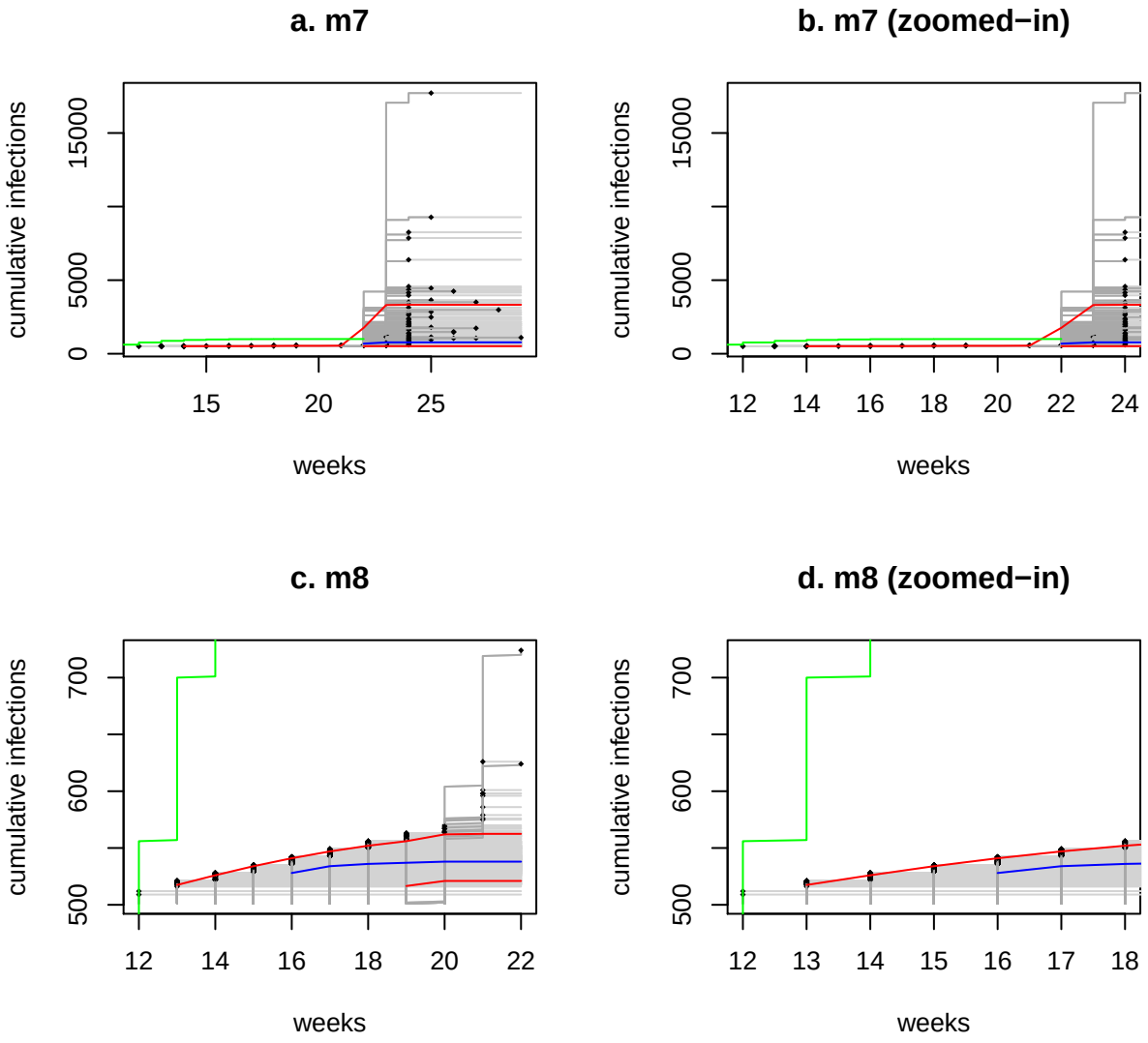


Figure 6.4. Forecasts for model 7 and 8: Poisson GLM applied to binned data. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast. The left panels (a and c) show the full length of the forecasts, whereas the right panels (panels b and d) show the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.

Model 5: weighted gamma GLM applied to binned data

Forecasts made with the weighted gamma GLM created a very broad forecasting interval (519 ; 4452) and contained the original epidemic. The median was close to the original epidemic with a value of 880. Although it is not easily visible in **Figure 6.3a** and **b**, only 270 forecasts were plotted, on which will be elaborated in the next subsection. The duration of the original epidemic was shorter than most of the forecasted epidemics, although the majority of bootstrapped forecast durations was within 23 weeks (161 days), which was near the original duration of approximately 150 days. The outliers in epidemic size were beneath the 90th quantile for duration. The shape of the forecasts made with the weighted gamma GLM was, like with model 3 and 4, contrary to the expectations: where a very steep curve was expected from c_{halt} onwards, turning into a declining growth, the epidemic seemed to start all over again with an exponentially increasing growth at first.

Model 6: weighted gamma GLM applied to underreported, binned data

Forecasts made with the weighted gamma GLM for underreported data generally underreported the original epidemic, which can be seen in **Figure 6.3c** and **d**. The forecasting interval did not contain the original epidemic. The original epidemic can be seen in the plot, where it was increasing rapidly at this point, whereas the forecasts showed a declining and slow growth in comparison. It is also interesting to see that the longest duration (18 weeks) and the 90th duration quantile (17 weeks) were close to each other, and that the shape of these forecasts did depict declining growth: models 3, 4, and 5 started with increasing growth and later turned into declining growth.

Model 7: Poisson GLM applied to binned data

It can be seen in **Figure 6.4a** and **b** that the forecasts made with the Poisson model on fully reported data creates a broad forecasting interval of (515; 3327), although the median of 0 seemed to be close to the original epidemic. A visible outlier grows to an epidemic size of beyond 15.000. The duration of the original epidemic was shorter than most of the forecasts, although the difference was moderate. The shape of the forecasts made with the fully reported Poisson GLM was exponentially increasing, where a declining growth was expected, as was mentioned at the subsection of model 5.

Model 8: Poisson GLM applied to underreported, binned data

It can be seen that the forecasts made with the underreported Poisson GLM underestimated the original epidemic. The forecasting interval (521 ; 562) did not contain the original epidemic, which can be seen to grow rapidly again like in model 6, and the median of 538 was far from the original epidemic. The duration of the original epidemic was slightly longer than most of the forecasts, and the highest epidemic size estimates were in the 90+ quantiles. The shape of the underreported Poisson forecasts from c_{halt} onwards was declining growth, although the strong growth that was expected at this point, was absent.

Conclusion

The forecasting results show a *systematic underreporting for all models in terms of epidemic size. Additionally, the shape of many forecasts was different than expected from the hypothesized model. Only the gamma GLM applied to individual level data was able to capture the intended shape from c_{halt} onwards. The other forecasts continued as if the epidemic was at the start with one infected individual again, starting with an exponentially growing curve, finishing with an exponentially declining growth. The models sometimes caught the intended declining growth from 50% of the epidemic onwards: this was generally the case for underreported models. However, the underreported models did generally not catch the quick growth at the point c_{halt}.

Table 6.2. Skipped bootstrap runs per cause for models 1-8.

	N NA	N = 25x	N <= 0	b NA	b <= 0
M1: default	0	0	0	0	0
M2: + underrep	0	0	0	0	0
M3: bin + jit	0	0	0	0	0
M4: + underrep	0	0	0	0	3
M5: bin + weight	225	1	0	0	0
M6: + underrep	4	0	0	0	0
M7: bin+pois	205	1	0	0	0
M8: + underrep	18	0	0	0	0

The duration of the forecasted epidemics was inconclusive with regard to the original epidemic, which is probably related to the shape issues described. The correctly shaped forecasts tended to overestimate the duration. There was no pattern for the underreported data. Models applied to (jittered) binned data did have the tendency to approximate the eventual duration.

From the simulation results, the gamma GLM applied to fully reported data seemed most promising in terms of the epidemic size, however the duration was overestimated. The weighted gamma GLM and Poisson GLM applied to fully reported data had better results in terms of duration and also contained the original epidemic within the forecasting intervals, even though the forecasting intervals were very broad. Contrary to the expectations the gamma GLM applied to jittered, fully reported data performed worse for the epidemic size than forecasts made with other models applied to fully reported data, where it was expected that this model would give similar results to the gamma GLM applied to individual data.

Skipped bootstrap runs

As mentioned earlier, a forecast was not made when $\hat{N}^*$ was larger than 25 times the original value of $N = 1000$ for the sake of computation time. This might have influenced the medians and forecasting intervals, as such it was checked how many bootstrap runs were skipped. **Table 6.2** and **6.3** show how many bootstraps were skipped for each model, divided over several reasons. `N_na`, `beta_na` and `both_na` tell us that the bootstrap run was skipped because either $\hat{\beta}^*$, $\hat{N}^*$ or both were `NA`. The tables show that data analyzed in binned format (without altering the data such as with jitter) produced many `NA` values for $\hat{N}^*$ after calculating back from the regression coefficients. `N_large` indicated that a run was skipped because the sample size was over 25 times the original epidemic size. This only seemed to happen incidentally, and only with the binned models. `beta_0` and `N_0` indicate that a run was skipped because either $\hat{\beta}^*$, $\hat{N}^*$, or both were zero or less than zero. This mainly seemed to happen with binned models, and mostly with the back-calculated values from the Poisson model. Three times the underreported gamma GLM applied to jittered data yielded an estimate of $\hat{\beta}^*$ equal to or below zero. Incidentally for the bin level models, $\hat{\beta}^*$ was equal to or below zero simultaneously with an $\hat{N}^*$ of `NA`.

To summarize, $\hat{N}^*$ becoming `NA` after calculating back seems to be the main reason to skip a bootstrap run, where it was expected that very large sample sizes would be the main problem. Especially with fully reported, binned data -analyzed with either a weighted gamma GLM or a Poisson GLM- almost many bootstraps were skipped. Strange to see is that the amount of skipped bootstrap runs of models applied to underreported, binned data sets was smaller than fully reported data sets.

A simple and easy solution for the high rate of `NA`'s in the binned models would be to enlarge the number of bootstrap samples: suppose half gets skipped, could we take double the bootstrap samples. As mentioned in Section 5, the exact collinearity introduced by the systematic model might be solved by introducing a different systematic model. Control statements for extremely large back-calculated values of $\hat{N}^*$ are advised: although rare, the estimate can peak into millions.

Table 6.3. Skipped bootstrap runs per cause for models 1-8 (continued).

	both NA	b NA & N > 25x	b NA & N <= 0	b <= 0 & N NA	b <= 0 & N > 25x	both 0
M1: default	0	0	0	0	0	0
M2: + underrep	0	0	0	0	0	0
M3: bin + jit	0	0	0	0	0	0
M4: + underrep	0	0	0	0	0	0
M5: bin + weight	0	0	0	4	0	0
M6: + underrep	0	0	0	4	0	0
M7: bin+pois	0	0	0	9	0	0
M8: + underrep	0	0	0	10	0	0

7. A case Study: Ebola

To illustrate the method, the Ebola data set that was made available by Backer and Wallinga (2016) was used. They applied a spatiotemporal model to the Ebola Epidemic in the three most heavily affected countries of the 2014 - 2016 Ebola epidemic: Guinea, Sierra Leone and Liberia. The data in this study was obtained from the WHO emergency response team (Aylward et al., 2014) and Gire et al. (2014), who made their data available to the research community during the Ebola epidemic.

The Ebola data consists of weekly incidence of Ebola cases in the patient database up to June 2015, as reported by the WHO. Information was reported for each district, such as country, prefecture, administrative centre, location coordinates, population size and observed weekly incidences.

In **Figure 7.1** and **Figure 7.2** the incidence and cumulative incidence per country are shown. It can be seen in **Figure 7.2** that the infections start growing exponentially from week 20 onward, and that around week 35 for Liberia and week 45 for Guinea and Sierra Leone, the infection declines and stops growing exponentially. The data set ends at 74 weeks: Guinea had infections recorded from week 1 until week 74, Liberia from week 12 until week 65, and Sierra Leone from week 21 until week 70. Starting with these plots as an indication, it would be safe to assume that the plotted shape of the simulations in **Figure 2.1** is similar to the plotted shape of the actual data. The epidemic size of the three countries is 11.317 for Sierra Leone, 4.994 for Liberia and 3.642 for Guinea. This is substantially larger than the epidemic size of $N = 1000$ that was used in the bootstrap procedure.

The Ebola data set is a weekly incidence data set. This indicates that it is a binned data set, which can be analyzed in several ways:

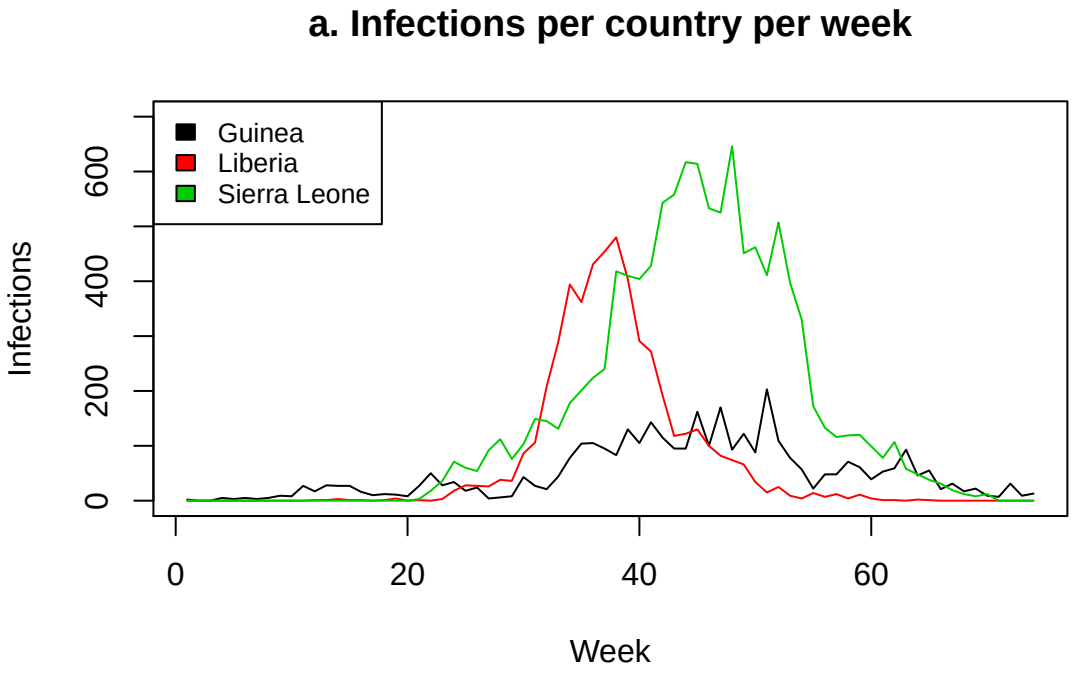
- the data could be jittered and analyzed with a gamma GLM;
- the data could be left as is and infection counts for each bin could be analyzed with a Poisson GLM;
- the data could be analyzed with a weighted gamma GLM, assigning weights to each bin equal to the number of infections in that week.

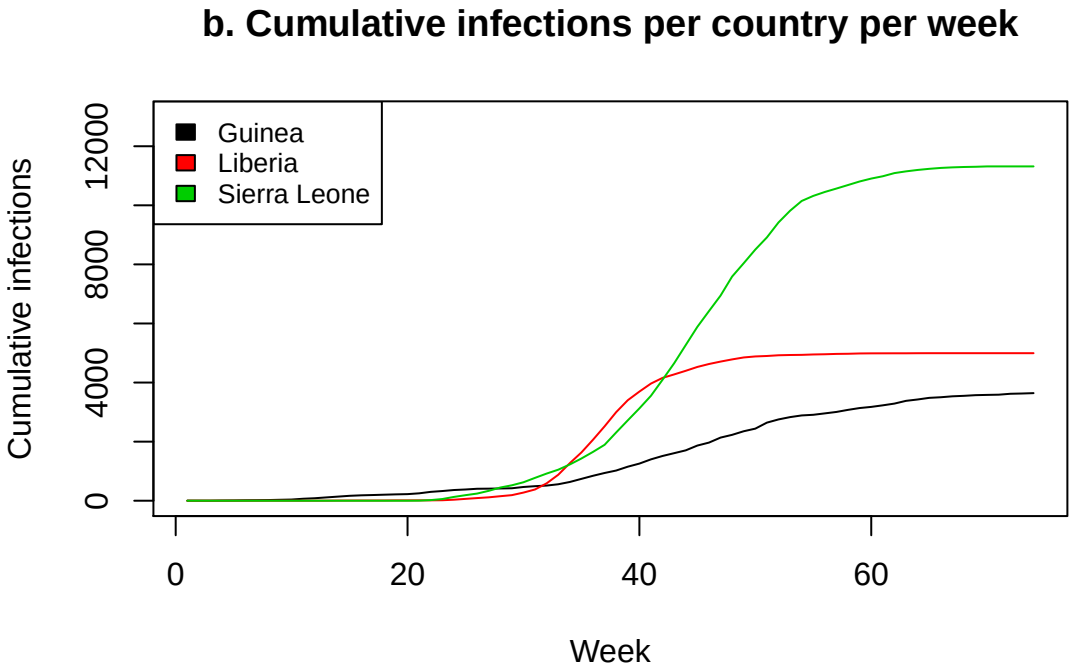
Of course, all three approaches mentioned above will be applied.

Key quantities

During the simulation it was assumed that an epidemic followed a specific shape, in which a parametric bootstrap was used to create new epidemics with estimates of β and N. A new infection would occur according to a Poisson process, where the time until the next event was given by an exponential distribution. Additionally, the distribution of the incubation time was assumed to be $Gamma(25, 0.4)$ (which is a mean of 10 days and a standard deviation of 2 days) and a specific reporting rate (either 1 or 0.25) was adopted.

To predict with and to make forecasts from real data, these assumptions should be checked or estimated from the data or from other possible sources. A reporting rate of Ebola could be derived from literature on the 2014 epidemic: Meltzer et al. (2014) used EbolaResponse to predict the number of beds in use on August





Incidence (a) and cumulative incidence (b) plot of Guinea (black line), Liberia(red line), and Sierra Leone (green line). The x-axis shows the number of weeks the epidemic is taking place, the y-axis shows the (cumulative) number of infections.

28, 2014. This estimate was compared to the actual number of beds in use, and an underreporting factor of 2.5 was calculated, corresponding to a reporting rate of 0.4. We believe that a reporting rate of 0.4 is realistic, because Ebola is a very severe disease with an average mortality rate of 50%, varying from 25% to 90% in past outbreaks (WHO, 2018). A proportion of the infected individuals might not make the health care facilities. Additionally, many issues were encountered in controlling the epidemic, such as shortage of health workforce, a lack of resources and a poor road network (Shoman, Karafillakis and Rawaf, 2017). However, for new infectious diseases, literature might not be easily available on the reporting rate.

Estimates of β and N could be calculated from the regression coefficients obtained by applying GLMs to the data at hand: by bootstrapping the data, bootstrap means and 90% bootstrap confidence intervals were calculated for the contact rate $\hat{\beta}$, epidemic size $\hat{N}$ and the regression coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. It was assumed that $\beta_1 \approx \beta$ when the sample size is sufficiently large. In the simulations all GLM models underestimated the actual epidemic size.

The general shape of the cumulative plots of the Ebola epidemics was similar to the simulated ones. Additionally, assuming that new infections would occur to a Poisson process with time to the next event given by an exponential distribution, similar values could be used as in the simulation. However, some values of the parameters associated with these distributions were to be determined, such as incubation time or reporting rate.

For incubation time literature could be used: the incubation time is a well-known factor for many infectious diseases. For Ebola, Aylward et al. (2014) tracked back the moment of infection by interviewing Ebola patients, determining a mean incubation time of 11.4 days. Therefore, the incubation time as in the simulations with a mean of 10 days and a standard deviation of 2 days, sufficed. However, research as was done by Aylward and colleagues is time-consuming and might not be feasible for new infectious diseases.

For new, unknown diseases, there is little information available for estimating the reporting rate and other parameters. A possible solution would be applying non-parametric bootstrapping to the data: the parameters that cannot be obtained from the literature, would hopefully be included automatically with this form of resampling data. For this section non-parametric bootstrapping was applied to the data.

Table 7.1. Epidemics and analysis methods used

Country	Data	Model
Guinea	binned, subsequently jittered	gamma GLM
Guinea	binned	weighted gamma GLM
Guinea	binned	Poisson GLM
Liberia	binned, subsequently jittered	gamma GLM
Liberia	binned	weighted gamma GLM
Liberia	binned	Poisson GLM
Sierra Leone	binned, subsequently jittered	gamma GLM
Sierra Leone	binned	weighted gamma GLM
Sierra Leone	binned	Poisson GLM

Methods

Three data sets were created from the 2014 Ebola data according to the three countries in the study: Guinea, Liberia and Sierra Leone. The three countries showed considerable variation in the cumulative incidence plots in **Figure 7.2**, of which Guinea was least similar to the simulated epidemics, and Liberia most similar. To compare the methods, all three data sets were analyzed with the three approaches available for binned data that were proposed in Section 3 on regression. An overview of the models and analysis approaches can be seen in **Table 7.1**.

To assess the predictions and to create forecasts, the data set was cut off at 50% (`psim=0.5`) and non-parametric bootstrapped predictions and forecasts were made. Our aim was to correctly forecast the eventual

size of the epidemic: the forecasted values of N were compared to the actual epidemics. The results of the bootstrapped predictions were checked to assess the quality of the models.

As making 500 forecasts proved to be extremely time-consuming, the approach was slightly different from the simulation set-up: 500 bootstraps were produced, which were used to create bootstrap statistics for the GLMs applied. Subsequently, the results of the first 100 bootstrap samples were used to create forecasts.

Results

Jittered data

The binned data was jittered to an individual level. **Figure 7.3** shows us the time-to-event plot of the jittered data sets for the three countries. Note the differing y-axes. It can be seen that the countries follow the previously seen bathtub pattern with long time-to-event in the tails and short time-to-event in the center. Guinea seemed to have slightly more variation in the tails. The patterns corresponded to the incidence plot in **Figure 7.1**, where Guinea had a less steep increase and less flat tails than the other two countries. Jittered data was analyzed with a gamma GLM.

Regression

GLMs were performed on the data of the three countries. For each analysis, a non-parametric bootstrap with 500 resamples was performed to obtain bootstrapped statistics of the regression coefficients (**Table 7.2**), of the estimates of the contact rate β and of the epidemic size N (**Table 7.3**). Fit statistics of the GLM were left out: if would be difficult to give the values a meaningful interpretation.

Regression coefficients

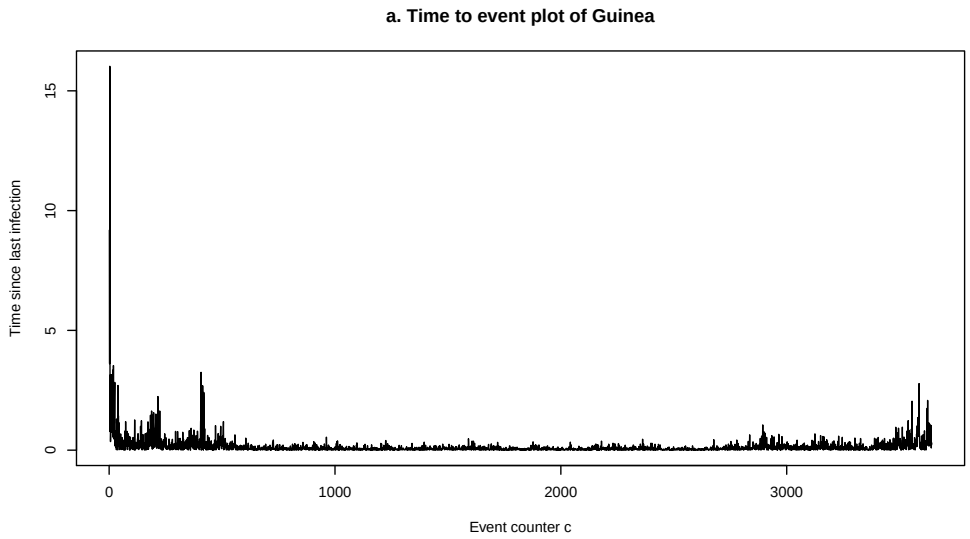
Table 7.2 shows the bootstrap regression coefficients β_0 for the intercept, β_1 for x and β_2 for x^2 of the GLMs applied. The bootstrap mean, standard deviation, 5th quantile and 95th quantile can be seen for Guinea, Liberia and Sierra Leone for the gamma, weighted gamma and Poisson GLM.

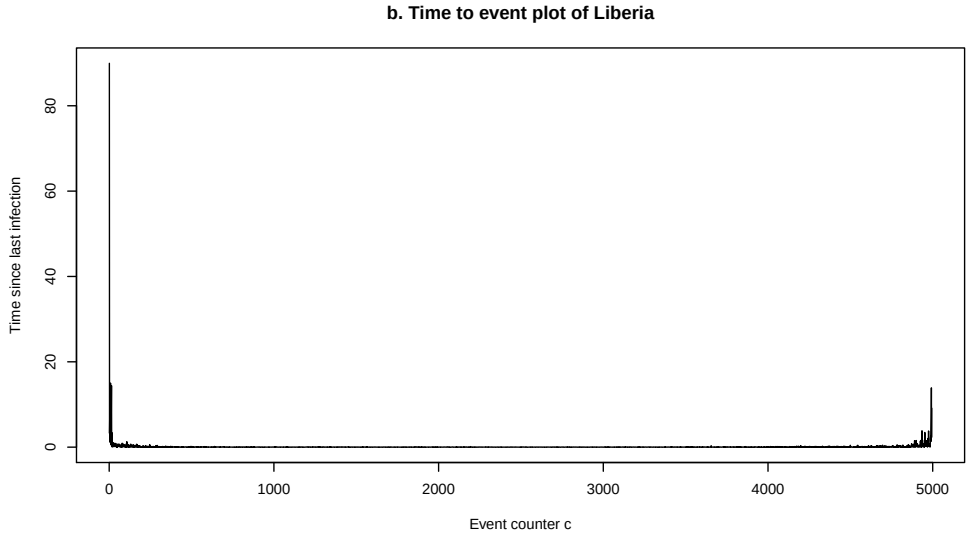
At a first glance the coefficients β_2 for variable x^2 were very small (<0.001) for both the gamma and Poisson GLM, which corresponded to the expectation that β_2 is generally a factor N smaller than β_1. The expected multiplication of 7 days for the coefficients (the bin level was week) was not clearly visible for either Poisson or weighted gamma GLM; the Poisson GLM did have some multiplication for the mean bootstrapped values, although irregular throughout each country and regression coefficient (β_0, β_1, or β_2).

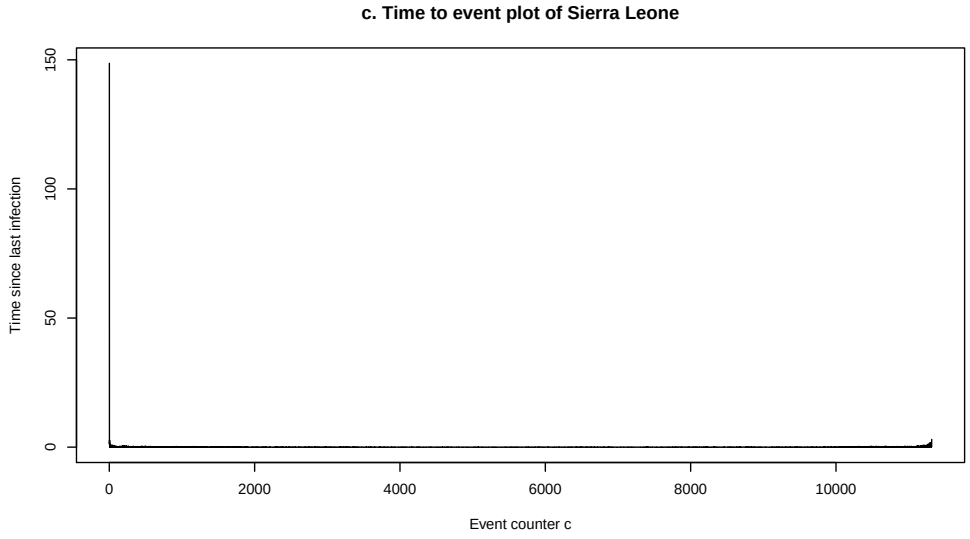
Almost none of the mean bootstrapped coefficients of the gamma and Poisson GLM had a 90% bootstrap confidence interval containing a zero, except for the Liberia intercepts of both methods, and the Sierra Leone intercept of the gamma GLM. Additionally, both Poisson and gamma GLM had a quite narrow confidence interval for β_1. All bootstrap confidence intervals for the weighted gamma GLM contained a zero; additionally the coefficients were much larger than of the other two analysis methods, indicating that the weighted gamma GLM might have some issues.

Comparing each analysis method across the three countries, it can be seen that the gamma GLM behaved consistently: the regression coefficients for x were not too large and for x^2 the coefficient was close to zero. When comparing the coefficients of the intercept to **Figure 7.2**, it could be said that for all three analysis methods, the intercept was credible: suppose this would be the actual starting count of the Poisson GLM, 22 infections would not have been extreme. Suppose the intercept of the (weighted) gamma GLM would represent the time between the first two infections, a starting value of 0.5 a day was also credible.

From the regression coefficients it was evident that none of the estimates of β_2 was explicitly negative, which was assumed in the calculations of Section 4. This might lead to estimation problems for the estimates of the contact rate and epidemic size, which will now be discussed.







Time-to-event plots of Guinea, Liberia and Sierra Leone. The x-axis shows the cumulative number of infections. The y-axis depicts the time between two subsequent infections.

Table 7.2. Bootstrapped regression coefficients

	Guinea				Liberia				Sierra Leone			
	mean	sd	5%	95%	mean	sd	5%	95%	mean	sd	5%	95%
gamma GLM												
intercept	0.539	0.206	0.127	0.952	-0.032	0.043	-0.118	0.053	1.452	1.902	-2.352	5.257
x	0.012	0.002	0.009	0.015	0.057	0.003	0.051	0.064	0.026	0.003	0.019	0.033
x2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
weighted gamma GLM												
intercept	0.516	0.266	-0.017	1.049	0.284	0.348	-0.411	0.980	3.390	5.221	-7.052	13.833
x	-0.186	0.122	-0.430	0.058	0.539	0.329	-0.119	1.196	-7.585	8.176	-23.937	8.768
x2	0.017	0.016	-0.016	0.050	-0.154	0.087	-0.328	0.019	4.932	2.937	-0.943	10.806
Poisson GLM												
intercept	4.836	0.679	3.479	6.194	0.193	0.240	-0.286	0.672	22.315	2.388	17.539	27.091
x	0.075	0.006	0.063	0.087	0.344	0.013	0.318	0.371	0.147	0.006	0.135	0.159
x2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 7.3. Bootstrapped estimates of Beta and N

	Beta					N				
	mean	sd	median	5%	95%	mean	sd	median	5%	95%
Guinea: N = 3642										
Gamma GLM	0.012	0.002	0.012	0.009	0.015	25601	294473	7442	-563346	614547
Weighted Gamma GLM	-0.171	0.106	-0.141	-0.383	0.042	95	1915	10	-3735	3925
Poisson GLM	0.075	0.006	0.075	0.063	0.087	-161126	7486194	-8743	-15133514	14811262
Liberia: N = 4994										
Gamma GLM	0.057	0.003	0.058	0.051	0.064	4677	492	4597	3693	5661
Weighted Gamma GLM	0.4	0.305	0.466	-0.21	1.011	3	13	4	-23	30
Poisson GLM	0.344	0.013	0.343	0.318	0.37	5594	586	5524	4423	6766
Sierra Leone: N = 11317										
Gamma GLM	0.026	0.003	0.028	0.019	0.033	15784	5856	12549	4071	27496
Weighted Gamma GLM	-3.523	4.718	-1.629	-12.959	5.913	2	0	1	1	2
Poisson GLM	0.147	0.006	0.147	0.135	0.159	24615	8093	22962	8430	40801

Contact Rate and Epidemic Size

From every set of coefficients, the contact rate β and the epidemic size N were estimated by using the formulae in Section 4. In **Table 7.3** bootstrapped statistics can be seen. The table is divided in two sets of five columns containing the bootstrapped mean, standard deviation, median, 5th quantile and 95th quantile: one set for the estimates of $\hat{\beta}$ and one set for $\hat{N}$.

Two issues can be spotted immediately when taking a first look at the table.

The first issue is that the weighted gamma GLM had no bootstrapped mean values of $\hat{\beta}$ and $\hat{N}$ without a zero in the 90% bootstrap confidence interval for any country. This implies that the estimates of the weighted gamma GLM experienced some issues. Looking at the actual values of $\hat{\beta}$ and $\hat{N}$ for the weighted gamma GLM, the bootstrapped mean of $\hat{\beta}$ was negative for Guinea and Sierra Leone and rather large for Liberia. The bootstrapped mean values of $\hat{N}$ were very small compared to the original epidemic value of $N = 4994$, or even negative.

The second issue concerns Guinea: all values of $\hat{N}$ seem somewhat off: all 90% bootstrap confidence intervals contained a zero, while $\hat{N}$ should not be able to drop into negative values. The confidence intervals for all three methods were also very large compared to the original value of $N = 3642$. It might be that the shape of Guinea in the cumulative incidence plot in **Figure 7.2** was not exponential enough to fit an exponential model on it.

Besides the two issues mentioned above, it seems that for both the gamma and Poisson GLM the estimated values $\hat{\beta}$ and $\hat{N}$ behaved well in terms of significance for Liberia and Sierra Leone, contrary to the weighted gamma GLM. It also stands out that for the mean values of $\hat{\beta}$, the values of the Poisson GLM were a factor 5-7 larger than the values of the gamma GLM. This is consistent with our expectations.

The median was included in this table, to assess whether the distribution of the estimates was skewed across the bootstrap samples. The median of $\hat{\beta}$ was generally close to the mean $\hat{\beta}$ for the gamma and Poisson GLM. The mean and median of $\hat{N}$ were also close for Liberia and Sierra Leone. As noted earlier, the values of $\hat{N}$ for Guinea were quite variable and will not be adressed.

There might be a pattern for the contact rate when comparing the estimates in **Table 7.3** to **Figure 7.2**. From the figure one would say that the contact rate would be incremental in size going from Guinea-Liberia-Sierra Leone, corresponding to the steepness of the increase halfway through the epidemics, or the size of the epidemic. However in **Table 7.3** the values of Liberia and Sierra Leone are reversed. It could be that, because Liberia has a single peak as can be seen in **Figure 7.1**, and Sierra Leone has a sort of plateau for some time, this comes back in the values of $\hat{\beta}$. The gamma GLM and Poisson GLM have the same pattern: the bootstrapped mean contact rate of the Poisson GLM is a bit less than seven times that of the gamma GLM. The weighted gamma GLM did not seem to have sensible results for the contact rate: negative values for Guinea and Sierra Leone, and none without a zero in the 90% bootstrap confidence intervals.

The estimated mean values $\hat{N}$ for the gamma and Poisson GLM in Liberia were quite close to the original value. The values of the gamma and Poisson GLM were respectively $N = 4677$ and $N = 5594$ compared to the original value of $N = 4994$. For Sierra Leone the Poisson GLM overestimated with a factor 2.2 ($N = 24615$), but the value of the gamma GLM was closer with a factor 1.4 ($N = 15784$) compared to the original value $N = 11317$. The Poisson GLM generally estimated a larger epidemic size than the gamma GLM: this can also be seen by looking at the 90% bootstrap confidence intervals. Looking at Guinea, none of the estimates of $\hat{N}$ was significant. The actual estimates show that an exponential model might not be the right model for Guinea.

The median $\hat{N}$ generally had a slightly lower value, but was close to the mean for Liberia and Sierra Leone. This outcome is nice as it indicates that the distributions of the bootstrapped values of N were not extremely skewed. For the median $\hat{\beta}$ this was true for the gamma and Poisson GLM of all three countries.

In general it can be said that the gamma and Poisson GLM performed quite well, where the weighted gamma GLM displayed some issues. The gamma GLM was closer to the original value of N for the largest epidemic -Sierra Leone- than the Poisson GLM, although still overestimating the eventual epidemic size. This was opposite compared to the simulation results of Section 5, where all methods underestimated the eventual epidemic size. It might be, looking at the results of Section 4, that the methods perform better with a larger epidemic size, as the epidemic size was 10.000 in Section 4, 1000 in Section 5, and 3642, 4994, and 11317 in Section 7.

Regarding the contact rate it is hard to say whether the results were realistic. The results seemed reliable for at least the gamma and Poisson GLM, although the contact rate seems a bit small, especially for a rapidly growing epidemic such as Ebola. It would be nice to compare this result with contact rates found in literature.

Forecasts

Earlier in this section the key quantities were discussed for the Ebola epidemics. Forecasts were made with the algorithm proposed in Section 6, with an incubation time with mean 10 and standard deviation 2 days. The rest of the key quantities was hopefully included by using a non-parametric bootstrap procedure. The 40% reporting rate mentioned earlier was taken into account while making forecasts. The forecasts in this subsection were made with the first 100 bootstrap GLM results of the previous subsection to limit the computing time. Because of the initial epidemic size of Sierra Leone being well over 10.000, the threshold for skipping a forecast due to a large estimate of N was set at 10 times the estimated value $\hat{N}$. A last remark is that, when the estimated number of infected individuals $\hat{N}$ was lower than c_{halt} at time point c_{halt}, which was halfway through the first estimated forecast, the algorithm would not run. This is logical, as the

Table 7.4. Skipped bootstrap runs for 100 bootstrap samples

	gamma GLM	weighted gamma GLM	Poisson GLM
Guinea	4	100	100
Liberia	0	100	0
Sierra Leone	0	100	0

Table 7.5. Skipped bootstrap runs for 100 bootstrap samples, per cause

	N NA	N = 10x	N <= 0	b NA	b <= 0	both NA	b NA & N > 10x	b NA & N <= 0	b <= 0 & N NA	b <= 0 & N > 10x	both 0
Guinea											
gamma GLM	0	4	0	0	0	0	0	0	0	0	0
weighted gamma GLM	0	2	0	0	0	0	0	0	0	98	0
Poisson GLM	0	100	0	0	0	0	0	0	0	0	0
Liberia											
gamma GLM	0	0	0	0	0	0	0	0	0	0	0
weighted gamma GLM	0	96	0	0	0	0	0	0	0	4	0
Poisson GLM	0	0	0	0	0	0	0	0	0	0	0
Sierra Leone											
gamma GLM	0	0	0	0	0	0	0	0	0	0	0
weighted gamma GLM	0	0	0	0	0	0	0	0	0	100	0
Poisson GLM	0	0	0	0	0	0	0	0	0	0	0

starting number of infected individuals would be higher than the end point $\hat{N}$. Therefore, on this occasion the estimated number of infected individuals $\hat{N}^*$ was restricted to be equal to c_{halt}, as this would be, in all cases, available information. In this case, the algorithm would run. It was not recorded how often the estimates $\hat{N}^*$ were restricted; in light of the prediction results, where a negative value of $\hat{N}^*$ was not unusual, this might be good to check.

Not all forecasts could be produced. **Table 7.4** shows the skipped number of bootstrap runs out of the desired 100 forecasts to be made, to which we now refer to as 'skipped forecasts'. **Table 7.5** shows the possible causes for a forecast to be skipped, divided by GLM type and country. The gamma GLM only produced a few skipped forecasts for all countries. The weighted gamma GLM produced 100 skipped forecasts for all countries. The Poisson GLM produced 100 skipped forecasts for the Guinea data, however, earlier it was established that Guinea did not follow the typically expected S-shape for the cumulative infections curve. It is possible that the Poisson GLM and the Guinea data do not combine in terms of the given systematical model or the identity link. No skipped forecasts were made for the Poisson GLM for Liberia and Sierra Leone. As mentioned in the previous paragraph, negative estimates $\hat{N}^*$ were restricted to be equal to c_{halt}, thus the number of skipped bootstrap runs due to negative estimates of $\hat{N}^*$ was zero. This restriction was mainly the case for Guinea, and incidentally for the weighted gamma GLM, as can be seen in the bootstrapped results for $\hat{N}^*$ in **Table 7.3**.

Looking at **Table 7.5** it can be seen that the skipped forecasts for the Guinea gamma and Poisson GLM were due to a large (>10x) epidemic size. For the Guinea weighted gamma GLM, mostly both the estimates of β were zero or negative and $\hat{N}$ was large. For Liberia and Sierra Leone no forecasts were skipped for the gamma and Poisson GLM, whereas all forecasts of the weighted gamma GLM were skipped, either because the estimate of N was large, or because the estimates of β were zero or negative.

The estimated regression coefficients are directly related to the negative values $\hat{\beta}$ and $\hat{N}$: when defining the systematic model and the estimation methods defined in Section 4, it was assumed that the regression coefficient β_1 of x had a positive value to enable the exponential growth, and the regression coefficient β_2 of x^2 to enable the declining growth of an epidemic. When both regression coefficients are either positive or negative, N becomes negative. As mentioned at the start of this subsection, these values of N were restricted to be positive. When $\hat{N}$ is sufficiently large, the estimate of β_1 equals β: is the regression coefficient negative, so is the estimate of the contact rate. With this in mind, the weighted gamma GLM might be very sensitive to the shape of the epidemic: the 'perfect' Liberia epidemic did produce regression estimates (no forecasts) for the weighted gamma GLM, although there were often not signi.

Table 7.8 shows the medians and 90% forecasting intervals of N, and the minimum, 90th percentile and

Table 7.8. Forecasting intervals of N, and minimum, 90th percentile and maximum duration from the cut-off point onward of the epidemic forecasts.

	N			duration		
	median	5%-ile	95%-ile	min	90%-ile	max
Guinea Gamma	13856	7392	100586	138	248	301
Liberia Gamma	6124	5507	6936	58	71	78
Sierra Leone Gamma	16240	14427	30864	92	137	154
Liberia Poisson	4535	3702	5591	9	11	12
Sierra Leone Poisson	21419	15800	32891	16	21	23

maximum duration in days or weeks. Remembering that the original epidemic size of Guinea was 3642, it can be seen that the median estimate of the Gamma forecasts overestimates this value with a factor 3.8 (2-27.6). The estimate of Liberia was closer, overestimating the original epidemic size of 4994 with a factor 1.2 (1.1-1.4). For Sierra Leone the original value was 11317, and also here the Gamma forecasts overestimated the epidemic size with a factor 1.4 (1.3-2.7). The Poisson GLM performed better for Liberia, with a median estimate of 4535, underestimating the original value with a factor 0.9 (0.7-1.1), but containing the original value in the forecasting interval. The original epidemics data ended at 74 weeks. Looking at the duration of the epidemics it can be seen that the gamma GLM applied to jittered data was measured in days, and the Poisson GLM in weeks. Important to note is that the duration in this table is from the time point corresponding to c_{halt} onwards, while the x-axes of the forecasts figures that follow measure from day 1 of the epidemic. For example, when c_{halt} is at thirteen weeks, and the end point of a specific epidemic is at 30 weeks, the end point will show up as 17 weeks in the table, and as 30 weeks in the figure.

Figure 7.4, 7.5, and **7.6** show forecasts for the gamma GLM for Guinea, Liberia and Sierra Leone, and **Figure 7.7** and **7.8** show forecasts for the Poisson GLM for Liberia and Sierra Leone. The forecasts are depicted in light grey, dark grey, and black dots. The dark grey lines are the original forecasts with their individual duration, with a black dot marking the end. The light grey lines are the forecasts that were stretched until the latest point of all forecasts together. The forecasting median is shown in blue; the 5th and 95th quantiles in red; and the original epidemic in green. On the x-axis the bin unit is given such as day or week, counting from the first moment of the epidemic, and on the y-axis the cumulative amount of infected individuals is shown, from the point c_{halt} onwards, although the starting point might be a bit difficult to see when the epidemic size of the forecasts grows large.

On each left panel, the full forecasts can be seen from the cut-off point until the latest forecasting duration, with the x-axis counting from day one: on each right panel, a 'zoomed-in' plot of the first 90 percent of epidemic durations is shown. As mentioned in Section 6, the epidemics can be incidentally stretched very far in time. The right-panel plots give information about whether the duration of the forecasts approximates the duration of the actual epidemics.

It can be seen from the figures that both gamma GLM and Poisson GLM systematically overestimate the eventual epidemic size, except for the Poisson GLM applied to the Liberia data. This is contrary to the results of the simulations, where all methods underestimated the original epidemic. Also contrary to the simulations was that many forecasted epidemics started with a rapid growth, continuing into a declining growth. However, the gamma GLM forecasts for Guinea and the Poisson forecasts for Sierra Leone started over again as if the epidemic starts at one infected individual, like in the simulations.

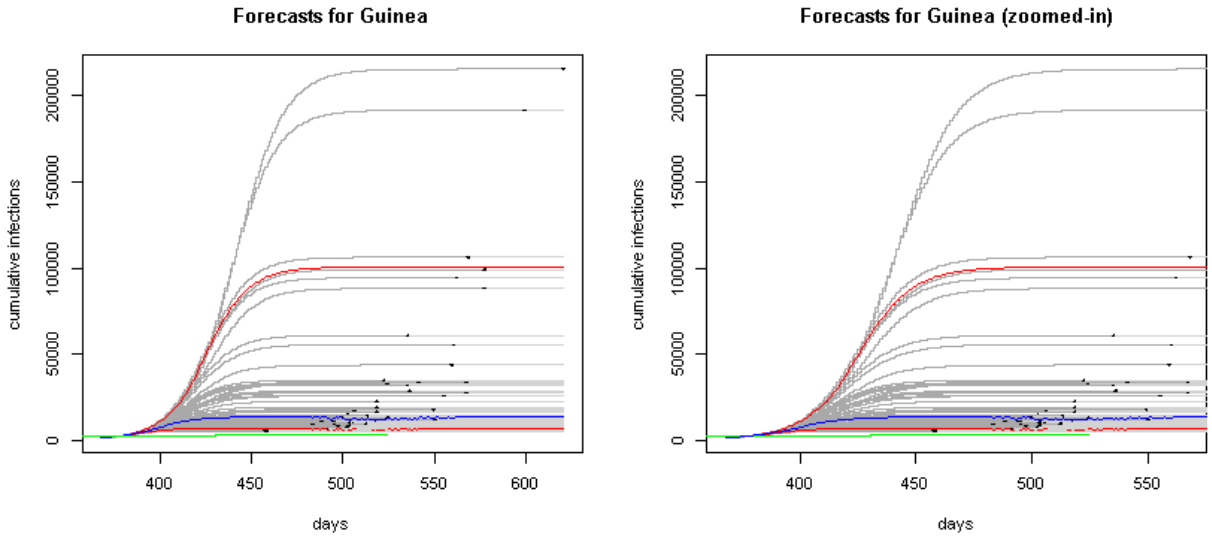
Regarding the duration of the epidemics: the data set shows records of 74 weeks, however Liberia and Sierra Leone do not have records on all of these weeks. It is debatable whether the duration of the whole data set should be used, the actual length of an epidemic is from the moment the first case is recorded until the lasts case is recorded.

Guinea had infections recorded from week 1 until week 74. For the Guinea gamma forecasts the end point was around 620 days (89 weeks), with the majority ending before approximately 575 days (82 weeks). In **Figure 7.4** it is evident that many epidemics have an end point near the end point of the original epidemic.

Liberia had records from week 12 until week 65, making the actual epidemic 53 weeks in length. The Liberia gamma forecasts ended at 343 days (49 weeks) with the majority ending before 48 weeks. Forecasts made with the Poisson GLM ended sooner: for Liberia the longest duration was 34 weeks, with the majority ending before 33 weeks. Thus, the duration of the Liberia epidemic was generally underestimated, even though it is nice to see that the duration was consistently estimated for both methods.

Sierra Leone had records from week 21 until week 70, making the actual duration 49 weeks. The Sierra Leone gamma GLM ended at approximately 475 days (68 weeks) with the majority ending before 66 weeks. The Poisson forecasts for Sierra Leone the longest duration was 47 weeks with the majority ending before 45. Here the methods are not corresponding with each other: the gamma forecasts either overestimate the duration, and the Poisson forecasts slightly underestimate the duration when the actual epidemic length is used; however, the gamma forecasts slightly underestimate the epidemic length, and the Poisson forecasts estimate gravely underestimate the duration when all 74 weeks are used for interpretation.

Generally it can be said the the duration of the gamma forecasts is longer than the duration of the Poisson forecasts. Both methods are consistent in the projected duration of the epidemics. When a forecasts projects a large epidemic size, the duration is generally longer.



Forecasts for Guinea made with the gamma GLM, with on the x-axis time in days and on the y-axis the cumulative number of infections. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panel shows the full length of the forecasts, whereas the right panel shows the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.

8. Discussion

Summary

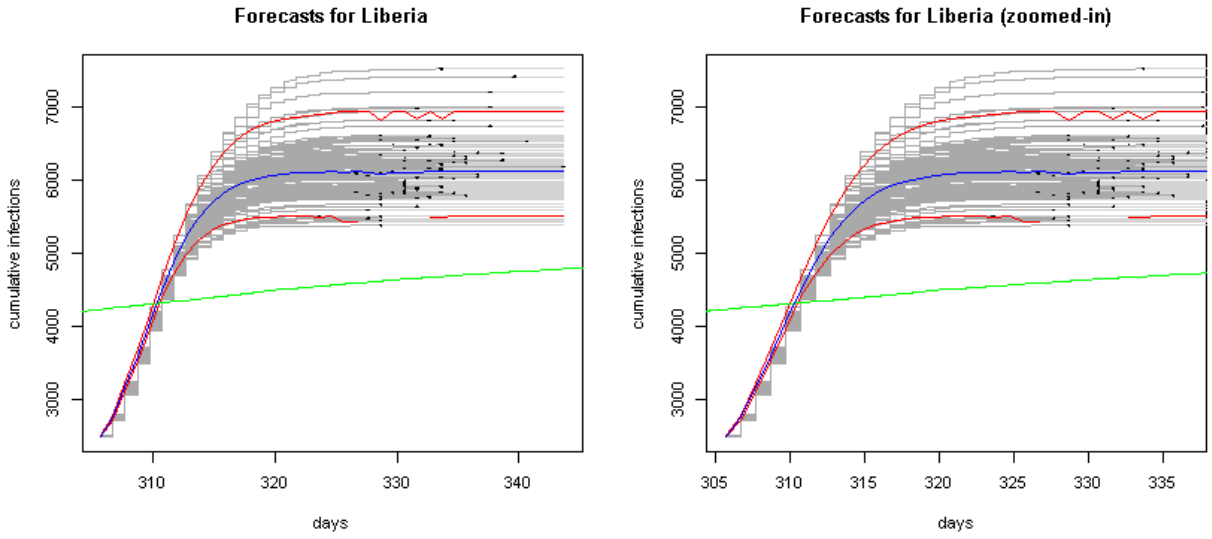
In this project a new approach to forecasting infectious disease epidemics was tested in a simulation and applied to data of the 2014 - 2016 Ebola epidemic. GLMs were applied to the (simulated) data, from which the key quantities contact rate and epidemic size could be obtained. With (non-)parametric bootstrapping, the GLM results could be assessed, and the key quantities were obtained and subsequently used to produce forecasts. Forecasting intervals were made to show the accuracy of the forecasts in terms of epidemic size and duration.

Simulation results

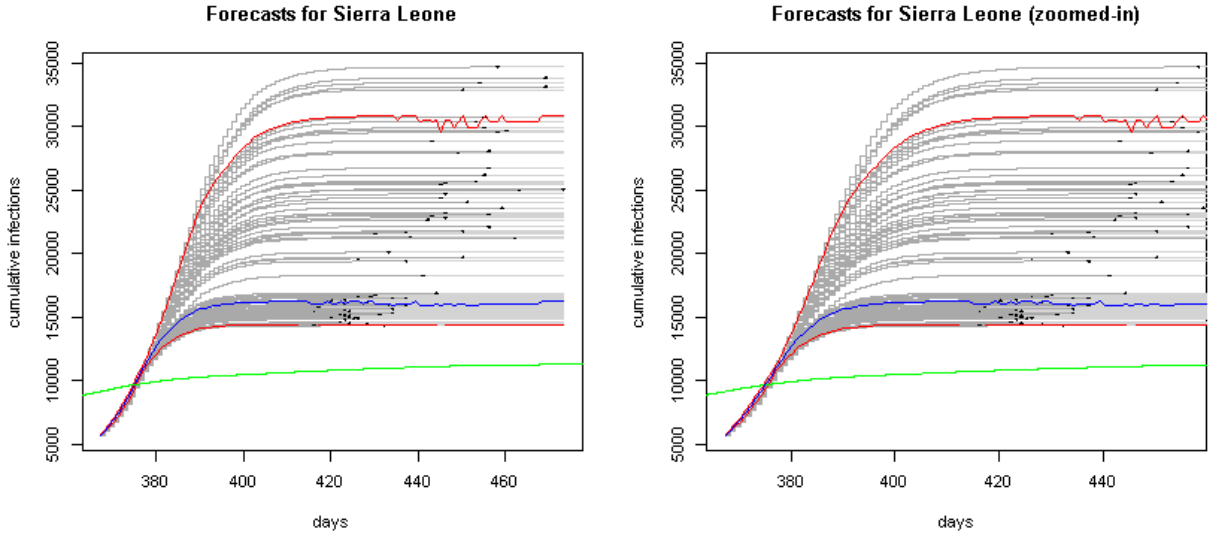
Results from the simulations suggested that the eventual epidemic size was underestimated, and that the contact rate was overestimated. Predictions from the gamma GLM were most accurate regarding both contact rate and epidemic size when compared to the original values of $\beta = 0.1$ and $N = 1000$.

Only GLMs applied to underreported data were able to consistently capture the intended negative value of the regression coefficient β_2 for x^2.

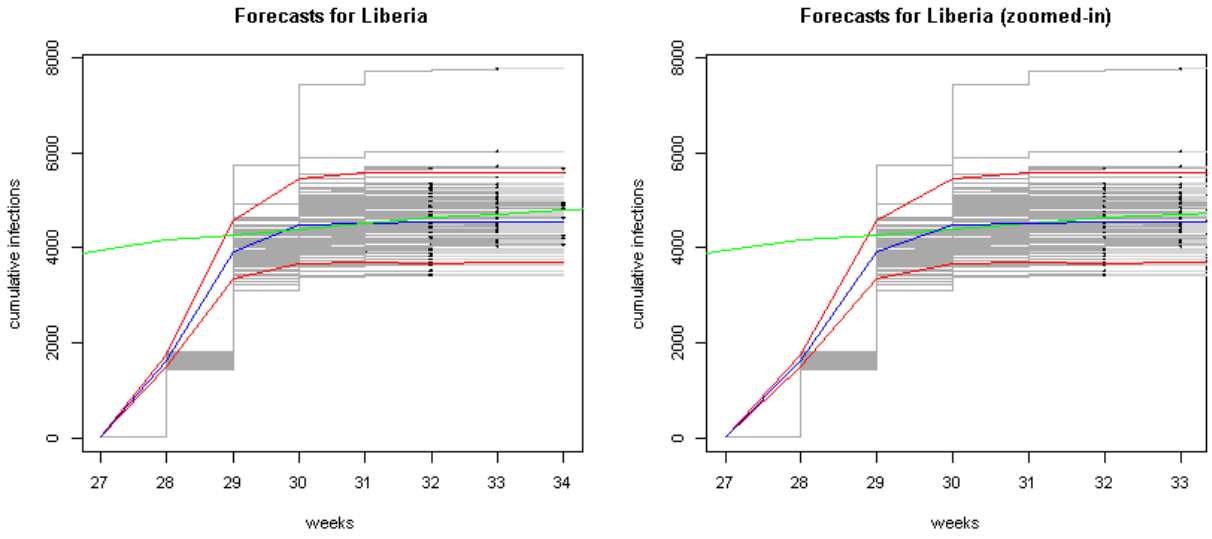
The Poisson GLM could use some improvements for estimating the contact rate, as it overestimated the original value with a factor 30: further investigation might be worthwhile, as the Poisson GLM is often used for analyzing epidemics in current research. The weighted gamma GLM had many issues and produced unreliable prediction results for the simulated data. The most issues were due to exact collinearity.



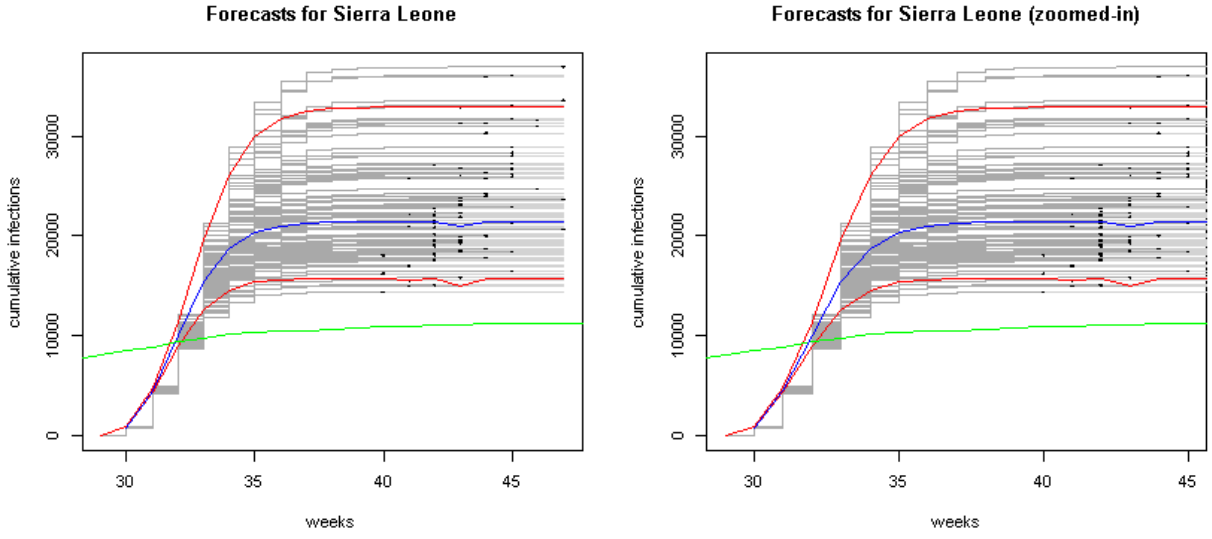
Forecasts for Liberia made with the gamma GLM, with on the x-axis time in days and on the y-axis the cumulative number of infections. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panel shows the full length of the forecasts, whereas the right panel shows the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.



Forecasts for Sierra Leone made with the gamma GLM, with on the x-axis time in days and on the y-axis the cumulative number of infections. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panel shows the full length of the forecasts, whereas the right panel shows the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.



Forecasts for Liberia made with the Poisson GLM, with on the x-axis time in weeks and on the y-axis the cumulative number of infections. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panel shows the full length of the forecasts, whereas the right panel shows the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.



Forecasts for Sierra Leone made with the Poisson GLM, with on the x-axis time in weeks and on the y-axis the cumulative number of infections. The light grey lines show the forecasts; the dark grey lines show the original duration of each forecast, of which the end point is marked with a black dot. The left panel shows the full length of the forecasts, whereas the right panel shows the forecasts up until the 90th duration quantile. The green line (if visible) is the original epidemic, the red lines show the 90% forecasting interval for the number of infected, and the blue line is the median number of infected individuals.

Ebola data application results

Applying the method to the 2014 Ebola data suggested that the epidemic size was overestimated. The majority of issues was due to extreme estimates of the epidemic size: forecasts for the weighted gamma GLM could not be made due to extremely large estimates of $\hat{N}$ or negative estimates of $\hat{\beta}$. The same was the case for the Poisson GLM for Guinea. For Liberia, the bootstrapped median estimate of the epidemic size was close to the original value when half of the data was used to fit the model for both the gamma and Poisson GLM. For Sierra Leone the epidemic size was overestimated slightly by the bootstrapped median of the gamma GLM, and with a factor 2 for the bootstrapped median of the Poisson GLM.

Regarding the results for the contact rate a comparison should be made to existing literature. The hypothesized multiplication of 7 days for the bootstrapped mean regression coefficient was present between the gamma GLM and the Poisson GLM, which is a nice result.

The gamma GLM tended to overestimate the epidemic duration, where the Poisson GLM generally underestimated. The duration was positively related to the size of the forecasted epidemic.

Issues

Prediction. The most important cause of issues with prediction and subsequently forecasting, was the systematic model. The model $y = \beta_0 + \beta_1 x + \beta_2 x^2$ was chosen to calculate back estimates of β and N. To calculate back these values, the regression coefficient β_1 for x should be positive to introduce exponential growth, and β_2 for x^2 negative to enter a phase of decreasing growth. The first issue arose when the coefficients were both positive or both negative: the estimation process of Section 4 would produce negative estimates of N. This could for example happen if the cumulative incidence data did not have the hypothesized shape. An easy, on-the-go solution for this issue would be to restrict negative estimates of N to be positive: however, negative estimates are a result of the mismatch between the shape of the data and the model. Just restricting estimates to have a positive sign will completely ignore this mismatch, which is the actual problem. Dropping the run for which the negative estimate of N was made, will result in throwing away information about the performance of the model. Additionally, less forecasts can be made as a result. It might be an idea to replace negative estimates with the mean of all (credible) estimates, or some other value that makes sense, such as historical epidemic size estimates of similar diseases.

The second issue due to the systematic model was that fitting the model on bin level introduced exact collinearity in the simulations. The GLM dropped the seemingly redundant predictor x^2, as it was linearly related to x. Without the regression coefficient β_2 for x^2 the estimates of N could not be made. In the simulations almost 50% of the forecasts could not be made for the weighted gamma GLM and the Poisson GLM, due to dropped parameters because of exact collinearity. It is probable that the number of observations, or rows, in the binned data sets was too small compared to the number of infected individuals in the counts in the x-variable: the issue did not arise with data analyzed on individual level: neither with jittered data nor with individually simulated data. This issue did not arise when applying the method to the Ebola data. A possible solution for this issue would be to orthogonalise the data with the `poly()`command in R. This would solve the collinearity issue, however, the interpretability of the coefficients would be lost. It was chosen not to use this solution, as the values and interpretation of the coefficients were important for comparison with the original values used to produce the simulations.

Another possible cause of issues with the Poisson GLM is the identity link that was used to compare the regression coefficients of the Poisson GLM with those of the gamma GLM. A solution might be using the log link, which is the default for the Poisson GLM. The log transforms large values to be drastically smaller, possibly solving the exact collinearity and other estimation issues. Calculating back the regression coefficients for comparing them to the gamma GLM would be an extra action, but it would definitely be worth it if the Poisson GLM would not encounter issues.

The weighted gamma GLM did not perform well overall. Besides the aforementioned issues, the use of weights might have complicated the estimation process: the weighted structure may not work well with the simplicity

of the systematic model; or the weighted structure is not the right approach with large counts (populations). While using the log link for a Poisson GLM is evident as a solution, it would be a challenge to use it here: for comparability to the gamma GLM we would have to do a log and an inverse transformation.

A final issue with the GLM was one that sporadically occurred in the simulations, but very often in the Ebola case: the estimated value of the epidemic size, $\hat{N}$ could grow extremely large. This occurred only once for the Poisson GLM in all simulations. It is advised to check for this issue, because a single outlier can drastically increase computation time.

Forecasting. Three main issues were present for the forecasts: as mentioned before, not all estimates of β and N could be made due to estimation problems of the GLM, which prevented the forecasts to run. The second issue was the systematic under- or overestimation of the eventual epidemic size in the simulations, where the results of the simulations were opposite to the results of the application to the Ebola data. The third issue concerns the hypothesized shape of the epidemic: because the epidemic was cut off halfway, the epidemic should enter a state of decreasing growth from the cut-off point onwards. Occasionally, the epidemic would "re-start" as if there was only one infected individual, starting with exponential growth again. It seems as if the model did not always capture the hypothesized shape.

When one has little information to base predictions and forecasts on, the uncertainty increases. The range of the forecasts was sometimes extremely large, even when made with half of the data. What will happen when the cut-off point is even sooner in the epidemic?

Improvements

For this project, a snowball effect was in place: when the GLM did not capture the shape of the data correctly, estimates of the coefficients, subsequently of the epidemic size and contact rate, and subsequently the forecasts, would go wrong. A major improvement would be finding a GLM that could determine or confirm the current state of an epidemic: are we at the start (exponential growth), somewhere in the center (slowing growth), beyond the peak (decreasing growth), or at the final stage (slow growth) of an epidemic? This could even be determined by visually checking the cumulative growth curve of an epidemic, even though the result would be a subjective one. Based on the state, the calculations for estimating the epidemic size and contact rate could be adapted with some expectation, or rules.

Bayesian analysis

Besides adapting the calculating back part of the method based on some classification, it might also be possible to put some restrictions on the calculation of the regression coefficients: restricting β_1 to be positive and β_2 to be negative would possibly prevent estimates of N and β to be unrealistic. It was chosen not to restrict the regression coefficients, as this would have involved adjusting results in our case, which is bad practise. Suppose the estimate of N was negative, and we would simply remove the negative sign, this would bring some unforeseen bias into the results. Additionally, without adjusting 'wrong' estimates, the size of the problem could be assessed. A possibility would be to replace the estimate with a mean value.

If a more complicated approach is feasible, applying a Bayesian analysis gives the possibility to include these constraints by means of the 'prior'. The constraints might cause the Bayesian algorithm to converge to a solution that has more realistic results than those obtained in this thesis, especially for the Poisson GLM and weighted gamma GLM. Additionally, possible 'human judgement' (Farrow et al., 2017) as was mentioned in the introduction, might supply valuable insights that can be included in a Bayesian analysis.

Another reason to try out a Bayesian approach is because it can handle complicated data: Wood (2010) stated that ecological systems are quite chaotic, making it difficult to reproduce actual epidemics and to obtain valid fit statistics in the field of dynamic models. Wood proposes a solution in which an MCMC chain is used, where summary statistics of the data are used to simulate mean and covariance matrices.

Other improvements

When using the individual level gamma GLM, it might be worthwhile to (if possible) take in account the reporting patterns: many health care instances are closed during the weekends. For example, when jittering a binned data set, one could specify a distribution that peaks on Monday and Friday. Other possible information to take into account would be seasonal data, environmental data such as the climate of an area, and as Backer and Wallinga (2016) did, taking into account spatio-temporal information, although the latter might not be feasible in a GLM setting. The aim of this study was to make a simple model and assess its performance in terms of model fit and forecasting accuracy; however, this approach might be too simple when other information is to be taken into account. A slightly more complicated method might be a better approach in such a case.

A possible improvement for this approach would involve assessing and testing several systematic models to see if there is another, sufficiently simple model that can be used in this setting. In this project a very simple model was used, which allowed us to easily estimate β and N. However, a slightly more complicated model would prevent the observed exact collinearity while improving the model fit on the data, hence possibly improving the forecasts as well. For the Poisson GLM it would be interesting to see if using a log link solves some of the colinearity problems.

Not only recorded information can be used to make improvements: a challenge that is related to the pattern of which day of the week people visit a doctor, would be to find out if there is a pattern in underreported data: which people do not end up in the system, and why? Related to possible patterns in the absence of recordings, would be to add more realistic scenarios to the simulations, such as people dying, getting better, or being or becoming immune.

A final possible improvement would be in one of the modeling choices: it was chosen to use a specific number of infected individuals as cut-off point. The alternative was to use some predetermined point in time, regardless of the number of infected individuals. For a simulation study it made sense to use the number of infected individuals: we decided the eventual number beforehand, and we knew how it was simulated. However, in a real epidemic disease outbreak, both the eventual number of infected individuals and the eventual duration are unkown. In such a case, it might be a better approach to use some predefined point in time from which to make forecasts.

Constraints

The major practical drawback of this study was the working memory of my laptop. Although this has challenged me into finding more efficient and resourceful solutions for analysis and programming, it could not prevent some simulations to run for hours.

Modeling constraint

The goal of this approach wat to create a simple approach. The simplicity can in many ways be a gain, however, it imposes many constraints as well. The simplicity of the chosen model might miss crucial systematic variation in the data, causing unreliable estimates of key quantities and the epidemic size, as was seen in the results. Another simplicity issue might arise due to the model with which the forecasts were made. Not all possible key quantities have been taken into account, and the relatively simple forecasts might not catch the actual epidemic.

Future implications

The eventual aim is to prevent infectious disease outbreaks and minimize the damage done to society. Improving forecasts closely relates to improvements on registration and detection of infectious diseases, implying more collaboration between health care instances: epidemics can be contained sooner if information

is quickly available on the spreading rate and size of the epidemic. One of the historically most effective measures to controlling infectious diseases has been vaccination (Khabbaz et al., 2014). Proper forecasts would enable quick decision-making for starting vaccination programs or deploying health care solutions. Additionally, having a relatively simple approach (GLMs applied to time-to-event data) available for the general scientific public might improve the collaboration and quick decision-making during epidemics. Current analysis models for epidemics are often complicated and specialized. It is very valuable if the severity of a new epidemic can be estimated quickly.

This project has shown that, with a relatively simple model, the size of an epidemic can be predicted reasonably well when only half of the epidemic has taken place, given that the shape of the epidemic on hand is somewhat exponential. It would be worthwile to investigate this approach further with some of the improvements mentioned earlier in this section, such as the log link for the Poisson GLM and improving the systematic model. For future research we recommend looking into a Bayesian approach for setting constraints on the regression parameters β_1 and β_2.

9. References

Aylward, B., Barboza, P., Bawo, L., Bertherat, E., Bilivogui, P., Blake, I., et al. (2014) Ebola virus disease in West Africa: the first 9 months of the epidemic and forward projections. *N Engl J Med., 371(16)*: 1481-1495. doi: https://doi.org/10.1056/NEJMoa1411100 PMID: 25244186

Backer, J. A., Wallinga, J. (2016). Spatiotemporal Analysis of the 2014 Ebola Epidemic in West Africa. *PLoS Comput Biol 12(12)*: e1005210. DOI: http://dx.doi.org/10.1371/journal.pcbi.1005210

Farrow, D. C., Brooks, L. C., Hyun, S., Tibshirani, R. J., Burke, D. S. & Rosenfeld, R. (2017) A human judgment approach to epidemiological forecasting. *PLoS Comput Biol 13(3)*: e1005248. https://doi.org/10.1371/journal.pcbi.1005248

Friesema, I. H. M., Koppeschaar C. E., Donker G. A., Dijkstra F., Van Noort S. P. & Smallenburg R., et al. (2009). Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in The Netherlands. *Vaccine, 27(45)*: 6353 - 7.

Gandon, S., Day, T., Metcalf, C. J. E. & Grenfell, B. T. (2016). Forecasting Epidemiological and Evolutionary Dynamics of Infectious Diseases. *Trends in Ecology & Evolution, 31(10)*: 776-788. http://dx.doi.org/10.1016/j.tree.2016.07.010

Gire, S. K., Goba, A., Andersen. K. G., Sealfon, R. S., Park, D. J., Kanneh, L., et al. (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science, 345(6202)*: 1369-1372. doi: https://doi.org/10.1126/science.1259657/ PMID: 25214632

Khabbaz, R. F., Moseley, R. R., Steiner, R. J., Levitt, A. M. & Bell, B. P. (2014). Challenges of infectious diseases in the USA. *The Lancet, 384(9937)*, 53-63. https://doi.org/10.1016/S0140-6736(14)60890-4

King, A., Ionides, E. L., Pascual, M. & Bouma, M. J. (2008) Inap- parent infections and cholera dynamics. *Nature, 454*: 877{880. DOI: http://dx.doi.org/10.1038/nature07084

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall.

McCullagh, P. (1992). Introduction to Nelder and Wedderburn (1972) Generalized Linear Models. In: Kotz, S., Johnson N. L. (eds) *Breakthroughs in Statistics.* Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY

Meltzer, M. I., Atkins, C. Y., Santibanez, S., Knust, B., Petersen, B. W., Ervin, E. D., Nichol, S. T., Damon, I. K., Washington, M. L. (2014). Estimating the Future Number of Cases in the Ebola Epidemic - Liberia and Sierra Leone, 2014-2015. *MMWR, 63(Suppl-3)*: 1-14.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, series A (General), 135(3)*: 370-384.

Ong, J. B. S., Chen, M. I.-C., Cook, A. R., Lee, H. C., Lee, V. J., Lin, R. T. P., Tambyah, P. A., Goh, L. G. (2010). Real-Time Epidemic Monitoring and Forecasting of H1N1-2009 Using Influenza-Like Illness from General Practice and Family Doctor Clinics in Singapore. *PLoS ONE 5(4)*: e10036.

https://doi.org/10.1371/journal.pone.0010036

Shoman, H., Karafillakis, E. & Rawaf, S. (2017). The link between the West African Ebola outbreak and health systems in Guinea, Liberia and Sierra Leone: a systematic review. *Globalization and Health, 13*: 1. https://doi.org/10.1186/s12992-016-0224-2 under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L. & Vespignani. A, (2017). The RAPIDD Ebola Forecasting Challenge: Synthesis and Lessons Learnt. *Epidemics* http://dx.doi.org/10.1016/j.epidem.2017.08.002

Wood, S. N. (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature, 466*: 1102-1104. DOI: http://dx.doi.org/10.1038/nature09319.

World Health Organization. (2018). Ebola Virus Disease. http://www.who.int/mediacentre/factsheets/fs103/en/